Almacenamiento de datos en AWS

Marzo de 2016

Avisos

Este documento se ofrece solo con fines informativos. Representa la oferta de productos y las prácticas actuales de AWS a partir de la fecha de publicación de este documento. Los productos y prácticas pueden modificarse sin previo aviso. Los clientes son responsables de realizar sus propias evaluaciones independientes de la información contenida en este documento y de cualquier uso de los productos o servicios de AWS, cada uno de los cuales se ofrece "como es", sin garantía de ningún tipo, ya sea explícita o implícita. Este documento no genera ninguna garantía, representaciones, compromisos contractuales, condiciones ni garantías de AWS, sus filiales, proveedores ni licenciantes. Las responsabilidades y obligaciones de AWS con respecto a sus clientes se controlan mediante los acuerdos de AWS, y este documento no forma parte ni modifica ningún acuerdo entre AWS y sus clientes.

Contenido

Resumen

Los ingenieros en datos, analistas de datos y desarrolladores en compañías de todo el mundo buscan migrar datos de almacenamiento a la cloud para aumentar el desempeño y reducir los costos. En este documento técnico se trata un enfoque moderno del análisis y la arquitectura de almacenamiento de datos, se describen los servicios disponibles en Amazon Web Services (AWS) para implementar esta arquitectura y se proporcionan patrones de diseño comunes para desarrollar soluciones de almacenamiento de datos mediante el uso de dichos servicios.

Introducción

En el mundo actual, los datos y el análisis son indispensables para los negocios. La mayoría de las compañías grandes han construido almacenes de datos con fines informativos y de análisis, mediante el uso de datos desde una variedad de recursos. Estos recursos incluyen sus propios sistemas de procesamiento de transacciones y otras bases de datos.

Sin embargo, construir y ejecutar un almacén de datos, que es un repositorio central de información proveniente de una o más fuentes, fue siempre complicado y costoso. La mayoría de los sistemas de almacenamiento de datos son complejos de configurar, cuestan millones de dólares en software inmediato y gastos de hardware, y pueden tomar meses de planificación, adquisición, implementación y utilización. Luego de haber realizado las inversiones iniciales y de haber configurado su almacén de datos, tiene que contratar un equipo de administradores de base de datos a fin de mantener un control rápido de consultas y protegerse contra la pérdida de datos.

Además, los almacenes de datos son difíciles de escalar. Cuando los volúmenes de datos aumentan o cuando desea realizar análisis o informes disponibles para más usuarios, tiene que elegir entre aceptar un desempeño de consultas lento o invertir tiempo y esfuerzo en un proceso de actualización costoso. De hecho, algunos equipos de TI desaconsejan aumentar los datos o agregar consultas, a fin

de proteger los acuerdos de nivel de servicios existentes. Muchas compañías se esfuerzan para mantener una buena relación con los proveedores de base de datos tradicionales. Generalmente, estas empresas están obligadas a actualizar el hardware para obtener un sistema administrado, o a entrar en un ciclo de negociación prolongado a fin de conseguir una licencia por un tiempo determinado. Cuando alcanzan el límite de escalado en un motor de almacenamiento de datos, están obligados a migrar a otro motor del mismo proveedor con una semántica de SQL diferente.

Amazon Redshift cambió la forma de pensar de las compañías sobre el almacenamiento de datos, al reducir significativamente el esfuerzo y los costos relacionados con la implementación de sistemas de almacenamiento de datos, sin comprometer las características y el desempeño. Amazon Redshift es una solución rápida y totalmente administrada de almacenamiento de datos de varios petabytes que permite analizar volúmenes grandes de datos de forma sencilla y económica, mediante el uso de las herramientas de inteligencia empresarial (BI) existentes. Con Amazon Redshift, puede obtener el desempeño de los motores de almacenamiento de datos en columnas que realizan massively parallel processing (MPP, procesamiento paralelo de forma masiva), a un décimo del costo. Puede comenzar a usarlo por 0.25 USD sin compromisos y ampliarlo después a varios petabytes, por 1,000 USD por terabyte anuales.

Desde su lanzamiento en febrero de 2013, Amazon Redshift ha sido uno de los servicios de AWS más rápidos en crecer, con miles de clientes en industrias y empresas de diferentes magnitudes. Compañías como NTT DOCOMO, FINRA, Johnson & Johnson, Hearst, Amgen y NASDAQ migraron a Amazon Redshift. Por esta razón, se calificó a Amazon Redshift como un líder en el informe *Forrester Wave: Informe Enterprise Data Warehouse, Q4 2015*.[1]

En este documento técnico, se proporciona la información necesaria para aprovechar el cambio estratégico que está sucediendo en el área de almacenamiento de datos desde las instalaciones hasta la cloud:

- Arquitectura de análisis moderna
- Opciones de tecnologías de almacenamiento de datos disponibles dentro de dicha arquitectura
- Una profundización en Amazon Redshift y sus características distintivas

- Un proyecto para desarrollar un sistema de almacenamiento de datos completo en AWS con Amazon Redshift y otros servicios
- Consejos prácticos para migrar desde otras soluciones de almacenamiento de datos y acceder a nuestro ecosistema de socios

Análisis moderno y arquitectura de almacenamiento de datos

Como ya se mencionó, un *almacén de datos* es un repositorio central de información proveniente de una o más fuentes. Generalmente, los datos fluyen a un almacén de datos desde sistemas de transacción y otras bases de datos relacionales, y normalmente incluyen datos estructurados, semiestructurados y datos sin estructurar. Estos datos se procesan, transforman e introducen a una cadencia regular. Los usuarios que incluyen científicos de datos, analistas de negocios y autoridades acceden a los datos por medio de las herramientas de BI, clientes de SQL y hojas de cálculo.

¿Para qué construir un almacén de datos? ¿por qué no se ejecutan las consultas de análisis directamente en una base de datos de procesamiento de transacciones online (OLTP, online transaction processing), donde se registran las transacciones? Para responder esta pregunta, observemos las diferencias entre almacenes de datos y bases de datos OLTP. Los almacenes de datos están optimizados para operaciones de escritura por lotes y para leer volúmenes grandes de datos, mientras que las bases de datos OLTP están optimizadas para operaciones de escritura continuas y volúmenes grandes de pequeñas operaciones de lectura. En general, los almacenamientos de datos utilizan esquemas no normalizados como el esquema Star y el esquema Snowflake debido a los altos requisitos de rendimiento, mientras que las bases de datos OLTP utilizan esquemas altamente normalizados, que son más adecuados para requisitos de rendimiento de transacción altos. El esquema Star consta de algunas tablas de hechos grandes que hacen referencia a una cantidad de tablas de dimensiones. El esquema Snowflake, una extensión del esquema Star, consta de tablas de dimensiones que se normalizan aun más.

Para aprovechar los beneficios de usar un almacén de datos administrado como un data store separado mediante su fuente OLTP u otro sistema de fuentes, le

recomendamos crear una canalización de datos eficaz. Esta canalización extrae los datos del sistema de fuentes, los convierte en un esquema adecuado para el almacenamiento de datos y luego los carga en un almacén de datos. En la siguiente sección, se analizan los componentes básicos de una canalización de análisis y los diferentes servicios de AWS que puede usar para diseñar la canalización.

Arquitectura de análisis

Las canalizaciones de análisis están diseñadas para controlar volúmenes grandes de flujos entrantes de datos, desde fuentes heterogéneas como bases de datos, aplicaciones y dispositivos.

Una canalización de análisis típica posee las siguientes etapas:

1. Recopilación de datos.
2. Almacenamiento de datos.
3. Procesamiento de datos.
4. Análisis y visualización de datos.

A modo de ejemplo, véase la Figura 1 a continuación.

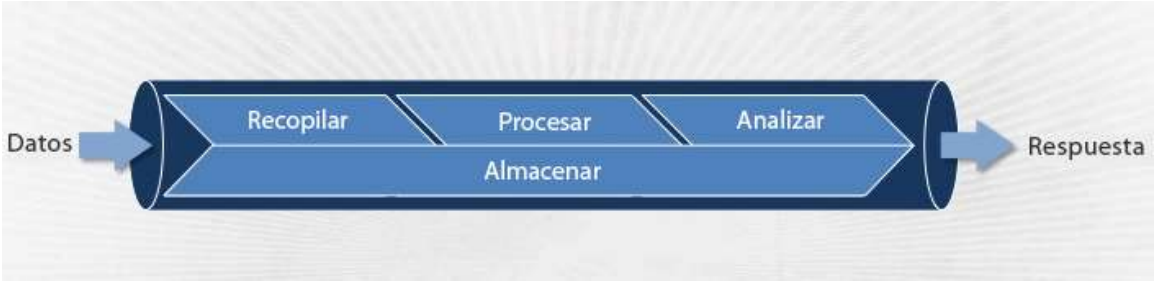

Figura 1: Canalización de análisis

Recopilación de datos

En la etapa de recopilación de datos, tenga en cuenta que probablemente posee diferentes tipos de datos, como datos de transacciones, datos del log, datos de streaming y datos de Internet de las cosas (IoT). Mediante AWS se proporcionan soluciones de almacenamiento de datos para cada uno de estos tipos de datos.

Datos de transacciones

Los datos de transacciones, como las transacciones de compras de comercio electrónico y transacciones financieras, generalmente se almacenan en sistemas de administración de bases de datos relacionales (RDBMS) o en sistemas de base de datos NoSQL. La elección de la solución de base de datos depende del caso de uso y de las características de la aplicación. Una base de datos NoSQL es idónea cuando los datos no están bien estructurados para ajustarse a un esquema definido o cuando el esquema cambia con mucha frecuencia. Por otra parte, una solución RDBMS es idónea cuando las transacciones suceden en múltiples filas de tabla y las consultas requieren combinaciones complejas. Amazon DynamoDB es un servicio de base de datos NoSQL completamente administrado que se puede usar como un almacén OLTP para sus aplicaciones. Amazon RDS le permite implementar una solución de base de datos relacional basada en SQL para su aplicación.

Datos del log

Los logs generados por un sistema de captura confiable lo ayudarán a solucionar problemas, llevar a cabo auditorías y realizar análisis mediante el uso de la información almacenada en los logs. Amazon Simple Storage Service (Amazon S3) es una solución de almacenamiento popular para datos no transaccionales, como los datos del log, que se usan para realizar análisis. Además, Amazon S3 es una solución de archivo popular, debido a que proporciona 11 nueves de durabilidad (es decir, 99,999999999 por ciento de durabilidad).

Datos de streaming

Las aplicaciones web, los dispositivos móviles y muchos servicios y aplicaciones de software pueden generar cantidades asombrosas de datos de streaming, algunas veces terabytes por hora, que se tienen que recopilar, almacenar y procesar continuamente.[2] Esto se puede realizar de forma sencilla y a un costo bajo mediante el uso de los servicios de Amazon Kinesis.

Datos del Internet de las cosas (IoT)

Los dispositivos y sensores de todo el mundo envían mensajes de manera continua. Las compañías se ven hoy frente a una necesidad en crecimiento de capturar estos datos y obtener inteligencia de ellos. Al usar IoT de AWS, los dispositivos conectados interactúan de forma fácil y segura con la cloud de AWS. El IoT de AWS facilita el uso de los servicios de AWS como AWS Lambda, Amazon Kinesis, Amazon S3, Amazon Machine Learning y Amazon DynamoDB

para crear aplicaciones que reúnen, procesan, analizan y actúan sobre los datos IoT, sin tener que administrar ninguna infraestructura.

Procesamiento de datos

El proceso de recopilación proporciona datos que pueden contener información útil. Usted puede analizar la información extraída para obtener la inteligencia que lo ayudará a hacer crecer su negocio. Esta inteligencia podría, por ejemplo, describir su comportamiento de usuario y la popularidad relativa de sus productos. La práctica recomendada para reunir esta inteligencia es cargar sus datos sin procesar en el almacenamiento de datos para realizar análisis adicionales.

Para realizar esto, existen dos tipos de flujos de trabajo de procesamiento, por lotes y en tiempo real. Las formas más comunes de procesamiento, procesamiento analítico online (OLAP) y OLTP, utilizan uno de estos tipos cada una. Generalmente, el procesamiento analítico online (OLAP) se basa en lotes. A diferencia del procesamiento OLAP, los sistemas OLTP están orientados hacia procesamientos en tiempo real y generalmente no son idóneos para procesamientos por lotes. Si desacopla el procesamiento de datos de su sistema OLTP, evita que el procesamiento de datos afecte su carga de trabajo de OLTP.

Primero se debe observar qué implica el procesamiento por lotes.

Extracción, transformación y carga (ETL)

ETL es el proceso de extracción de datos de múltiples fuentes para cargarlos en sistemas de almacenamiento de datos. Generalmente, ETL es un proceso continuo con un flujo de trabajo definido de forma precisa. Durante este proceso, los datos se extraen inicialmente desde una o más fuentes. Luego, los datos extraídos se limpian, enriquecen, transforman y cargan al almacén de datos. Las herramientas del marco de trabajo de Hadoop, como Apache Pig y Apache Hive, generalmente se usan en una canalización ETL para realizar transformaciones en volúmenes grandes de datos.

Extracción, carga y transformación (ELT)

ELT es una variante de ETL donde los datos extraídos se cargan primero al sistema de destino. Las transformaciones se realizan después de que se cargan los datos al almacén de datos. Normalmente, ELT trabaja de forma adecuada cuando su sistema de destino es lo suficientemente poderoso para gestionar las

transformaciones. Amazon Redshift se suele usar en canalizaciones ELT debido a que es altamente eficaz para realizar transformaciones.

Procesamiento analítico online (OLAP)

Los sistemas OLAP almacenan datos históricos acumulados en esquemas multidimensionales. Los sistemas OLAP, ampliamente utilizados en extracciones de datos, le permiten extraer datos y detectar tendencias en múltiples dimensiones. Amazon Redshift se usa con frecuencia para crear sistemas OLAP debido a que está optimizado para combinaciones rápidas.

Ahora veamos qué implica el procesamiento de datos en tiempo real.

Procesamiento en tiempo real

Anteriormente, se analizaron los datos de streaming y se mencionó a Amazon Kinesis como una solución para capturar y almacenar datos de streaming. Puede procesar estos datos de forma secuencial y creciente, sobre la base de registros o ventanas de tiempo variable, y usar los datos procesados para una gran variedad de análisis que incluyen correlaciones, agregaciones, filtrado y muestreo. Este tipo de procesamiento se denomina procesamiento en tiempo real. La información obtenida del procesamiento en tiempo real brinda a las compañías visibilidad de muchos aspectos de sus negocios y de las actividades de los clientes, como el uso del servicio (para realizar la medición o facturación), la actividad del servidor, los clics del sitio web y la geolocalización de dispositivos, personas y productos físicos. Esto permite a las compañías responder rápidamente a las situaciones emergentes. El procesamiento en tiempo real requiere una capa de procesamiento altamente concurrente y escalable.

Para procesar datos de streaming en tiempo real, puede usar AWS Lambda. Por medio de Lambda se pueden procesar los datos directamente de AWS IoT o Amazon Kinesis Streams. Lambda le permite ejecutar códigos sin aprovisionar o administrar servidores.

Amazon Kinesis Client Library (KCL, biblioteca de clientes de Amazon Kinesis) es otra forma de procesar datos de Amazon Kinesis Streams. KCL le permite obtener más flexibilidad que AWS Lambda para procesar sus datos entrantes por lotes a fin de realizar procesamientos adicionales. Además, puede usar KCL para aplicar transformaciones y personalizaciones en su lógica de procesamiento.

Amazon Kinesis Firehose es la forma más fácil de cargar datos de streaming en AWS. Puede capturar datos de streaming y cargarlos automáticamente en Amazon Redshift, lo que posibilita análisis casi en tiempo real mediante las herramientas de BI existentes y los paneles que ya está usando. Mediante Firehose, puede definir sus reglas de procesamiento por lotes y luego este se encarga de agrupar los datos y de proporcionárselos a Amazon Redshift de forma confiable.

Almacenamiento de datos

Puede almacenar sus datos en un almacén de datos o en un data mart, tal como se explica a continuación.

Almacén de datos

Como se expresó anteriormente, un *almacén de datos* es un repositorio central de información proveniente de una o más fuentes. Al usar almacenes de datos, puede ejecutar análisis rápidos sobre volúmenes grandes de datos y patrones descubiertos, que se encuentran ocultos en sus bases de datos, mediante las herramientas de BI. Los científicos de datos consultan un almacén de datos para realizar análisis sin conexión y detectar tendencias. Los usuarios de la organización consumen los datos mediante el uso de consultas ad hoc SQL, informes periódicos y paneles a fin de tomar decisiones de negocios críticas.

Data mart

Un *data mart* es una forma simple de almacenamiento de datos centrada en un área funcional o tema específicos. Por ejemplo, puede poseer data marts específicos para cada división en su organización o segmentar los data marts basándose en regiones. Puede construir data marts a partir de un almacén de datos grande, de almacenes operativos o de una combinación de ambos. Los data marts son fáciles de diseñar, construir y administrar. Sin embargo, ya que estos se centran en áreas funcionales específicas, la consulta en todas las áreas se puede transformar en una tarea compleja debido a la distribución.

Puede usar Amazon Redshift para construir data marts además de almacenes de datos.

Análisis y visualización

Luego de procesar los datos y ponerlos a disposición para realizar análisis adicionales, necesita las herramientas adecuadas a fin de analizar y visualizar los datos procesados.

En muchos casos, puede realizar el análisis de datos mediante el uso de las mismas herramientas que utiliza para procesar datos. Puede usar herramientas como SQL Workbench para analizar sus datos en Amazon Redshift con ANSI SQL. Amazon Redshift también funciona de forma adecuada con soluciones de BI de terceros disponibles en el mercado.

Amazon QuickSight es un servicio de BI, rápido y basado en la cloud que facilita la creación de visualizaciones, la realización de análisis ad hoc y la obtención rápida de perspectivas de los negocios a partir de sus datos. Este servicio está integrado con Amazon Redshift y actualmente se encuentra en versión preliminar, con disponibilidad general para fines del 2016.

Si está usando Amazon S3 como su principal forma de almacenamiento, una forma común de realizar el análisis y la visualización es ejecutar blocs de notas de Apache Spark en Amazon Elastic MapReduce (Amazon EMR). Al usar este proceso, se obtiene la flexibilidad para ejecutar SQL o un código personalizado escrito en lenguajes tales como Python y Scala.

Apache Zeppelin, otra estrategia de visualización, es una solución de BI de fuente abierta que se puede ejecutar en Amazon EMR para visualizar datos en Amazon S3 mediante el uso de Spark SQL. Además, puede usar Apache Zeppelin para visualizar datos en Amazon Redshift.

Canalización de análisis mediante servicios de AWS

AWS ofrece un amplio conjunto de servicios para implementar una plataforma de análisis integral. La Figura 2 muestra los servicios ya analizados y dónde se ajustan dentro de la canalización de análisis.

Figura 2: Canalización de análisis mediante servicios de AWS

Opciones de tecnologías de almacenamiento de datos

En esta sección, se analizan las opciones disponibles para crear un almacenamiento de datos: bases de datos orientadas a filas, bases de datos orientadas a columnas y arquitecturas de procesamiento paralelo de forma masiva.

Bases de datos orientadas a filas

Generalmente, las bases de datos orientadas a filas almacenan filas completas en un bloque físico. Se logra un alto desempeño en operaciones de lectura por medio de índices secundarios. Las bases de datos como Oracle Database Server, Microsoft SQL Server, MySQL y PostgreSQL son sistemas de bases de datos orientadas a filas. Estos sistemas se usaron tradicionalmente para el almacenamiento de datos, pero son más adecuados para procesamientos de transacciones online (OLTP) que para análisis.

Para optimizar el desempeño de un sistema basado en filas que se utiliza como un almacén de datos, los desarrolladores usan una serie de técnicas, incluida la creación de vistas materializadas, la creación de tablas de resumen de agregados previos, la construcción de índices en todas las combinaciones posibles de predicados, la implementación de la partición de datos para aprovechar el recorte de partición mediante el optimizador de consultas y las combinaciones basadas en el índice de desempeño.

Los almacenes de datos basados en filas tradicionales están limitados por los recursos disponibles en una sola máquina. Los data marts reducen el problema hasta cierto punto mediante el uso de sharding funcional. Puede dividir su almacén de datos en data marts múltiples y que cada uno satisfaga las necesidades de un área funcional específica. Sin embargo, cuando los data marts se agrandan con el paso del tiempo, el procesamiento de datos se vuelve más lento.

En un almacén de datos basado en filas, cada consulta tiene que leerse a través de todas las columnas de todas las filas de los bloques que satisfacen el predicado de la consulta, incluidas las columnas que no eligió. Este enfoque crea un cuello de botella significativo en el desempeño de los almacenes de datos, donde las tablas tienen más columnas, pero sus consultas usan solo unas pocas.

Bases de datos orientadas a columnas

Las bases de datos orientadas a columnas organizan cada columna en su propio conjunto de bloques físicos en lugar de embalar todas las filas en un bloque. Esta funcionalidad les permite ser más eficientes en E/S para consultas de solo lectura, ya que solo tienen que leer esas columnas a las que se accede mediante una consulta desde el disco (o desde la memoria). Este enfoque hace que las bases de datos orientadas a columnas sean una mejor opción que las bases de datos orientadas en filas para el almacenamiento de datos.

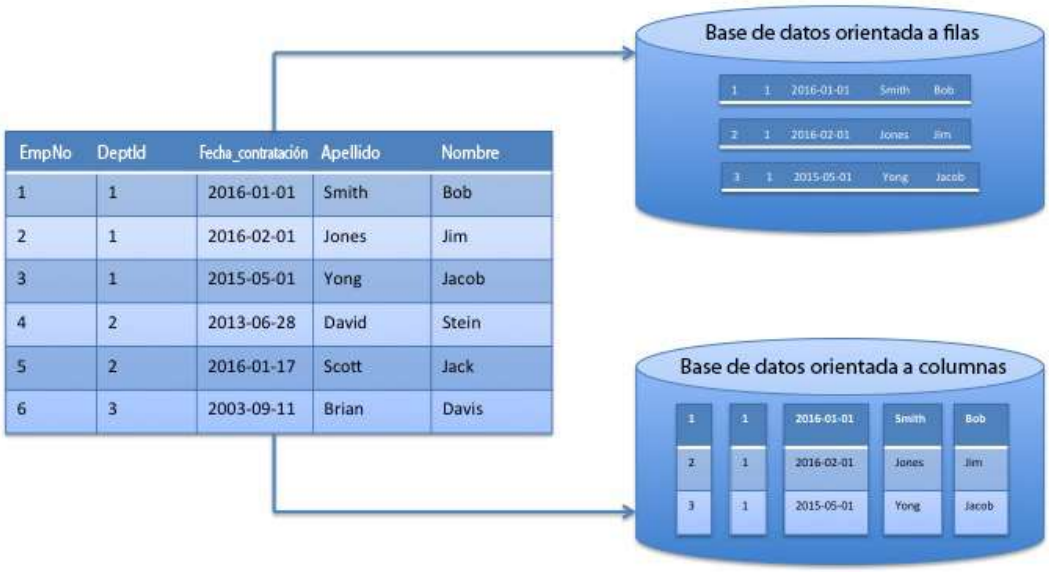

Figura 3: Bases de datos orientadas en fila en comparación con bases de datos orientadas a columnas

La Figura 3, anterior, ilustra la diferencia principal entre las bases de datos orientadas en fila y las bases de datos orientadas a columnas. Las filas están empaquetadas en sus propios bloques en una base de datos orientada en filas, y las columnas están empaquetadas en sus propios bloques en una base de datos orientada a columnas.

Después de una E/S más rápida, el mayor beneficio de usar una base de datos orientada a columnas es la compresión mejorada. Debido a que cada columna está empaquetada en su propio conjunto de bloques, cada bloque físico tiene el mismo tipo de datos. Cuando todos los datos son del mismo tipo, la base de datos puede usar algoritmos de compresión extremadamente eficientes. Como resultado, necesita menor almacenamiento en comparación con una base de datos orientada en filas. Este enfoque también significa gestionar una E/S significativamente menor, porque los mismos datos se almacenan en menos bloques.

Algunas bases de datos en columnas que se utilizan para el almacenamiento de datos son Amazon Redshift, Vertica, Teradata Aster y Druid.

Arquitecturas de procesamiento paralelo de forma masiva

Una arquitectura de procesamiento paralelo de forma masiva (MPP) le permite utilizar todos los recursos disponibles en el clúster para el procesamiento de los datos, lo cual aumenta significativamente el desempeño de los almacenes de datos de escala petabyte. Los almacenes de datos MPP le permiten mejorar el desempeño con solo agregar más nodos al clúster. Amazon Redshift, Druid, Vertica, GreenPlum y Teradata Aster son algunos de los almacenes de datos construidos con una arquitectura MPP. Marcos de trabajo de código abierto como Hadoop y Spark también admiten MPP.

Profundización Amazon Redshift

Como tecnología MPP en columnas, Amazon Redshift ofrece beneficios clave para el almacenamiento de datos rentable y efectivo, incluida la compresión eficiente, reducción de E/S y menores requisitos de almacenamiento. Se basa en ANSI SQL para que pueda ejecutar consultas existentes con poca o ninguna modificación. Como resultado, se ha convertido en una opción popular para los almacenes de datos empresariales y de data marts actuales. En esta sección, nos sumergimos más profundamente en Amazon Redshift y analizamos más sus capacidades.

Amazon Redshift ofrece un rápido desempeño de consulta y E/S rápido para prácticamente cualquier tamaño mediante la utilización del almacenamiento en

columnas y mediante la paralelización y distribución de las consultas en varios nodos. Automatiza la mayoría de las tareas administrativas comunes asociadas con el aprovisionamiento, la configuración, monitorización, realización de backup y protección de un almacén de datos, por lo que es muy fácil y barato de administrar. Mediante el uso de esta automatización puede crear un almacén de datos a escala de petabyte en solo unos minutos en lugar de las semanas o los meses que requiere una implementación local tradicional.

Desempeño

Amazon Redshift utiliza almacenamiento en columnas, compresión de datos y asignaciones de zona para reducir la cantidad de operaciones de E/S necesarias para realizar consultas. La clasificación intercalada permite un desempeño rápido sin la sobrecarga de mantener índices o proyecciones.

Amazon Redshift emplea una arquitectura MPP, que se beneficia de todos los recursos disponibles al paralelizar y distribuir operaciones SQL. El hardware subyacente está diseñado para un procesamiento de datos de alto desempeño, para lo que utiliza almacenamiento conectado local que maximiza el desempeño entre los CPU y las unidades, y una red de malla 10 GigE que maximiza el desempeño entre los nodos. El desempeño puede ajustarse en función de sus necesidades de almacenamiento de datos: AWS ofrece computación de alta densidad (Dense Compute, DC) con unidades de estado sólido y también opciones de almacenamiento de alta densidad (Dense Storage, DS). El despliegue continuo de actualizaciones de software ofrece mejoras de desempeño en curso sin intervención del usuario.

Durabilidad y disponibilidad

Para proporcionar la mejor durabilidad y disponibilidad de datos posible Amazon Redshift detecta y reemplaza automáticamente cualquier nodo defectuoso del clúster de almacén de datos. Habilita el nodo de sustitución de inmediato y carga los datos a los que se obtiene acceso con más frecuencia para que pueda reanudar la consulta de sus datos lo más rápido posible. Como Amazon Redshift refleja los datos de todo el clúster, usará los datos de otro nodo para reconstruir el nodo fallido. El clúster está en modo de solo lectura hasta que se aprovisiona y añade un nodo de sustitución al clúster, una acción que normalmente solo lleva unos minutos.

Los clústeres de Amazon Redshift residen en una sola zona de disponibilidad.[3] Sin embargo, si desea disponer de varias zonas de disponibilidad AZ para Amazon Redshift, puede crear una imagen reflejada y después administrar automáticamente la replicación y la conmutación por error.

Con solo unos clics en la consola de administración de Amazon Redshift puede configurar un entorno de recuperación de desastres (disaster recovery, DR) sólido con Amazon Redshift. Puede guardar copias de las copias de seguridad en múltiples regiones de AWS. En caso de una interrupción del servicio en una región AWS, puede restaurar el clúster desde la copia de seguridad en una región AWS diferente. Puede obtener acceso de lectura/escritura a su clúster a los pocos minutos de iniciar la operación de restauración.

Escalabilidad y elasticidad

Con tan solo unos clics en la consola o una llamada API, puede cambiar fácilmente el número y el tipo de nodos en el almacén de datos a medida que cambien sus necesidades de desempeño o capacidad.[4] Amazon Redshift le permite comenzar con solo un nodo de 160 GB y ampliarlo hasta un petabyte o más de datos de usuario comprimidos utilizando varios nodos. Para obtener más información, consulte Acerca de clústers y nodos en la *guía de administración de clústers de Amazon Redshift*.[5]

Cuando se cambia el tamaño, Amazon Redshift coloca el clúster existente en modo de solo lectura, aprovisiona un nuevo clúster del tamaño que desee y realiza una copia en el nuevo clúster de los datos del clúster anterior. Durante este proceso, usted solo paga el clúster de Amazon Redshift activo. Puede continuar realizando consultas en el clúster anterior mientras se aprovisiona el nuevo. Después de que los datos se hayan copiado al nuevo clúster, Amazon Redshift redirige las consultas automáticamente al nuevo clúster y elimina el clúster anterior.

Puede utilizar las acciones de la API de Amazon Redshift para lanzar clústeres mediante programación, escalar clústeres, crear copias de seguridad, restaurar copias de seguridad y mucho más. Al usar este enfoque, puede integrar estas acciones API en su stack de automatización existente o crear la automatización personalizada que se adapte a sus necesidades.

Interfaces

Amazon Redshift cuenta con conectividad de base de datos java (Java Database Connectivity, JDBC) y unidades de conectividad de base de datos abierta (Open Database Connectivity, ODBC) que puede descargar de la pestaña **Connect Client (Conexión del cliente)** de la consola, lo que significa que puede usar una amplia gama de clientes familiarizados con SQL. También puede usar unidades PostgreSQL JDBC y ODBC estándar. Para obtener más información sobre las unidades Amazon Redshift consulte Amazon Redshift and PostgreSQL en la *guía de desarrollador de base de datos de Amazon Redshift*. [6]

También se pueden encontrar numerosos ejemplos de integraciones validadas con muchos proveedores BI y ETL populares.[7] En estas integraciones, las cargas y descargas se ejecutan en paralelo en cada nodo de computación para maximizar la velocidad a la que se puede ingresar o exportar datos hacia y desde múltiples recursos, incluidos Amazon S3, Amazon EMR y Amazon DynamoDB. Puede cargar datos de streaming en Amazon Redshift con Amazon Kinesis Firehose, lo que posibilita análisis casi en tiempo real mediante las herramientas de BI y los paneles existentes. Puede encontrar las métricas de uso de computación, utilización de memoria, utilización de almacenamiento y tráfico de lectura/escritura del clúster de almacén de datos de Amazon Redshift a través de la consola o de las operaciones API CloudWatch deAmazon.

Seguridad

Para ayudar a proporcionar seguridad de los datos, puede ejecutar Amazon Redshift dentro de una cloud privada virtual basada en el servicio Amazon Virtual Private Cloud (Amazon VPC). Puede utilizar el modelo de redes definido por software de la VPC para definir reglas de firewall que restringen el tráfico basado en las reglas que usted configure.[8] Amazon Redshift admite conexiones habilitadas por SSL entre la aplicación cliente y su clúster de almacén de datos Amazon Redshift, lo que permite que los datos se cifren en tránsito.

Los nodos de computación Amazon Redshift almacenan sus datos, pero solo se puede acceder a ellos desde el nodo principal del clúster. Este aislamiento proporciona otra capa de seguridad. Amazon Redshift se integra con AWS CloudTrail para que pueda auditar las llamadas API de Amazon Redshift.[9] Para ayudar a mantener sus datos seguros en reposo, Amazon Redshift encripta cada bloque mediante el uso del cifrado AES-256 con aceleración por hardware a

medida que cada bloque se escribe en el disco. Este cifrado se lleva a cabo en un nivel bajo en el subsistema de E/S; el subsistema de E/S encripta todo lo escrito en el disco, incluidos los resultados intermedios de la consulta. Se realiza una copia de seguridad de los bloques como están, lo que significa que las copias de seguridad también están cifradas. De forma predeterminada, Amazon Redshift se encarga de administrar las claves, pero usted puede elegir administrar sus claves con sus propios módulos de seguridad de hardware (hardware security modules, HSM) o administrar sus claves a través del servicio de administración de claves de AWS.[10],[11]

Modelo de costos

Para usar Amazon Redshift no es necesario afrontar gastos anticipados ni asumir compromisos a largo plazo. Esto lo libera del enfoque de precios del gasto de capital y la complejidad de la planificación y compra de capacidad para el almacén de datos antes de que surja la necesidad. Los cargos se basan en el tamaño y el número de nodos en el clúster.

No se aplica ningún cargo adicional por el almacenamiento de copias de seguridad de hasta el 100% del almacenamiento aprovisionado. Por ejemplo, si dispone de un clúster activo con dos nodos XL para un total de 4 TB de almacenamiento, AWS le proporciona hasta 4 TB de almacenamiento de backup en Amazon S3 sin ningún costo adicional. El almacenamiento de backup más allá del tamaño del almacenamiento provisto y los backups almacenados después de terminar el clúster se facturan según las tarifas estándares de Amazon S3.[12] No hay cargo por transferencia de datos para la comunicación entre Amazon S3 y Amazon Redshift. Para obtener más información, consulte los precios de Amazon Redshift.[13]

Patrones de uso ideales

Amazon Redshift es ideal para el procesamiento analítico en línea (OLAP) mediante las herramientas BI existentes. Las organizaciones emplean Amazon Redshift para hacer lo siguiente:

- Ejecutar BI empresarial y presentar informes
- Analizar datos de ventas globales de varios productos
- Almacenar datos bursátiles históricos

- Analizar impresiones publicitarias y clics
- Acumular datos de juegos
- Analizar tendencias sociales
- Medir la calidad asistencial, la eficacia de las operaciones y el desempeño financiero en el área de atención sanitaria

Patrones de uso no recomendados

Amazon Redshift no es adecuado para los siguientes patrones de uso:

- **Pequeños conjuntos de datos** – Amazon Redshift se ha diseñado para el procesamiento en paralelo en un clúster. Si el conjunto de datos es inferior a 100 gigabytes, no va a obtener todos los beneficios que Amazon Redshift tiene para ofrecer y Amazon RDS puede ser una solución mejor.
- **OLTP** – Amazon Redshift se ha diseñado para cargas de trabajo de almacenamiento de datos con capacidades de análisis rápidas y económicas. Si lo que necesita es un sistema de transacciones rápido, un sistema de bases de datos relacionales tradicional basado en Amazon RDS o una base de datos de NoSQL como Amazon DynamoDB puede ser la mejor opción.
- **Datos no estructurados** – los datos de Amazon Redshift deben estar estructurados mediante un esquema definido. Amazon Redshift no es compatible con una estructura de esquema arbitrario para cada fila. Si sus datos no están estructurados, puede realizar la extracción, transformación y carga (ETL) en Amazon Elastic MapReduce (Amazon EMR) para preparar los datos para su carga en Amazon Redshift. Para datos JSON, puede almacenar pares de valor clave y usar las funciones nativas de JSON en sus consultas.[14]
- **Datos de objetos binarios grandes (BLOB)** – si piensa almacenar archivos de datos grandes binarios (BLOB) como video digital, imágenes o música, quizás le convenga almacenar los datos en Amazon S3 y hacer referencia a su ubicación en Amazon Redshift. En este escenario, Amazon Redshift realiza un seguimiento de los metadatos (como el nombre del elemento, el tamaño, la fecha de creación, el propietario la ubicación y demás) de los objetos binarios, pero los objetos grandes en sí estarían almacenados en Amazon S3.

Migración a Amazon Redshift

Si decide migrar de un almacén de datos existente a Amazon Redshift, la estrategia de migración a elegir depende de varios factores:

- El tamaño de la base de datos y sus tablas
- El ancho de banda de red entre el servidor de origen y AWS
- Si la migración y conversión al sistema AWS se llevará a cabo en un paso o una secuencia de pasos a lo largo del tiempo
- La tasa de cambio de datos en el sistema de origen
- Las transformaciones durante la migración
- La herramienta de colaboración que tiene previsto utilizar para la migración y ETL

Migración de un solo paso

La migración de un solo paso es una buena opción para bases de datos pequeñas que no requieren operación continua. Los clientes pueden extraer las bases de datos existentes en archivos de valores separados por comas (comma-separated value, CSV), luego utilizar servicios como AWS Import/Export Snowball para entregar bases de datos a Amazon S3 para cargar en Amazon Redshift. Después, los clientes prueban la base de datos de destino Amazon Redshift para la consistencia de los datos con la fuente. Una vez que han pasado todas las validaciones, la base de datos se conmuta a AWS.

Migración de dos pasos

La migración de dos pasos se utiliza comúnmente para las bases de datos de cualquier tamaño:

1. **Migración de datos inicial:** Los datos se extraen de la base de datos fuente, preferiblemente durante el uso no pico para minimizar el impacto. Luego, los datos se migran a Amazon Redshift mediante el seguimiento del enfoque de migración de una sola etapa descrito anteriormente.
2. **Migración de datos cambiados:** Los datos que cambiaron en la base de datos de origen después de la migración de datos inicial se propaga al

destino antes de la conversión. Este paso sincroniza las bases de datos de origen y de destino. Una vez que se hayan migrado todos los datos modificados, puede validar los datos en la base de datos de destino, realizar las pruebas necesarias y si estas se aprueban cambiarse al almacén de datos de Amazon Redshift.

Herramientas para la migración de la base de datos

Están disponibles muchas herramientas y tecnologías para la migración de datos. Puede usar algunas de estas herramientas de forma intercambiable o puede utilizar otras herramientas de terceros o de código abierto disponibles en el mercado.

1. El servicio de migración de base de datos de AWS admite los procesos de migración de uno y dos pasos descritos anteriormente.[15] Para seguir el proceso de migración de dos pasos, habilite el registro complementario para capturar los cambios en el sistema de origen. Puede habilitar el registro suplementario en el nivel de tabla o base de datos.
2. Las herramientas asociadas para la integración de datos adicionales son las siguientes:
    - Attunity
    - Informatica
    - SnapLogic
    - Talend
    - Bryte

Para obtener más información sobre la integración de datos y los socios consultores, véase Amazon Redshift Partners.[16]

Diseño de flujos de trabajo de almacenamiento de datos

En las secciones previas, se analizan las características de Amazon Redshift que lo hacen idóneo para el almacenamiento de datos. Para comprender cómo diseñar flujos de trabajo de almacenamiento de datos con Amazon Redshift, se debe observar el patrón de diseño más común junto con un ejemplo de caso de uso.

Suponga que una firma de confección multinacional tiene más de mil establecimientos comerciales, vende algunas líneas de ropa a través de grandes almacenes y tiendas de descuentos, y tiene presencia online. Desde una perspectiva técnica, estos tres canales operan actualmente de forma independiente. Poseen sistemas de punto de venta y de administración y departamentos contables diferentes. Ningún sistema combina todos los conjuntos de datos relacionados para proporcionarle al director una vista de 360 grados del negocio completo.

Suponga también que el director desea obtener una imagen completa de estos canales de la compañía y ser capaz de realizar análisis ad hoc tales como los siguientes:

- ¿Qué tendencias existen en los canales?
- ¿Qué regiones geográficas funcionan mejor con los distintos canales?
- ¿Son eficaces los anuncios y promociones de la compañía?
- ¿Qué tendencias existen en cada línea de ropa?
- ¿Qué factores externos, como las tasas de desempleo o las condiciones meteorológicas, afectan a las ventas de la compañía?
- ¿Cómo afectan a las ventas las características de las tiendas (por ejemplo, la permanencia de los empleados y la administración, centro comercial abierto versus centro comercial cerrado, la ubicación de los artículos en la tienda, las promociones, los artículos más visibles, las circulares de ventas, los expositores, etc.)?

Un almacén de datos para la compañía resuelve este problema. Recopila datos de cada uno de los sistemas de los tres canales y también de datos disponibles de forma pública, como los informes meteorológicos y económicos. Cada fuente de

datos envía datos diariamente para el sistema almacenamiento de datos. Debido a que cada fuente de datos puede estar estructurada de forma diferente, se realiza un proceso de extracción, transformación y carga (ETL) para reformatear los datos en una estructura común. Luego, podrán realizarse análisis de los datos de todas las fuentes simultáneamente. Para ello, se usa la siguiente arquitectura de flujo de datos:

Figura 4: Flujo de trabajo de almacenamiento de datos de la compañía

1. El primer paso de este proceso es ingresar los datos de fuentes diferentes a Amazon S3. Mediante Amazon S3, se proporciona una plataforma de almacenamiento de gran duración, económica y escalable en la que se puede escribir en paralelo desde muchas fuentes diferentes a un costo muy bajo.
2. Se usa Amazon EMR para transformar y limpiar los datos del formato de origen a fin de enviarlos al formato de destino. Amazon EMR está integrado con Amazon S3, que permite subprocesos de rendimiento paralelos desde cada nodo del clúster de Amazon EMR hacia y desde Amazon S3.

   Normalmente, un almacén de datos obtiene datos nuevos por las noches. Debido a que no se necesitan análisis a mitad de la noche, el único requisito de este proceso de transformación es que finalice por la mañana, cuando el director y otros usuarios profesionales deben acceder a informes y paneles. Por lo tanto, puede usar el mercado de subastas de Amazon EC2 para reducir aún más los costos de ETL en este caso.[17] Una buena estrategia de subasta es comenzar a pujar a un precio muy bajo a medianoche e ir aumentando el precio con el tiempo hasta que se consiga la capacidad. Cuando se aproxime la fecha límite, si las pujas no han tenido éxito, puede recurrir a los precios bajo demanda para asegurarse de que todavía satisface sus requisitos de plazo de finalización. Cada fuente puede tener un proceso de transformación diferente

en Amazon EMR. Sin embargo, mediante el modelo de pago por uso de AWS, puede crear un clúster de Amazon EMR separado para cada transformación y ajustarlo para que tenga la capacidad adecuada, y así completar todos los trabajos de transformación sin competir con los recursos de otros trabajos.

3. Cada trabajo de transformación carga los datos formateados y limpios en Amazon S3. Aquí se usa otra vez Amazon S3 porque Amazon Redshift puede cargar estos datos en paralelo desde Amazon S3, mediante el uso de múltiples subprocesos de cada nodo de clúster. Además, Amazon S3 proporciona un registro histórico y sirve como la fuente de confianza formateada entre sistemas. Otras herramientas pueden usar los datos ubicados en Amazon S3 para los análisis si se presentan nuevos requisitos con el tiempo.
4. Amazon Redshift carga, ordena, distribuye y comprime los datos en sus tablas para que puedan ejecutarse consultas analíticas eficazmente y en paralelo. A medida que aumente el tamaño de los datos con el paso del tiempo y se expanda el negocio, puede aumentar la capacidad al agregar más nodos.
5. Para visualizar los análisis, puede usar Amazon QuickSight o una de las muchas plataformas de visualización de los socios que se conectan con Amazon Redshift mediante el uso de controladores ODBC o JDBC. Aquí es donde el director y su personal ven informes, paneles y gráficos. Ahora, los directivos pueden usar los datos para tomar mejores decisiones sobre los recursos de la compañía, lo que en última instancia aumenta los ingresos y el valor para los accionistas.

Usted puede ampliar esta arquitectura flexible de manera sencilla cuando se expande su negocio, se abren canales nuevos, se lanzan nuevas aplicaciones móviles específicas de los clientes y se introducen más fuentes de datos. No implica más que hacer unos pocos clics en la consola de administración de Amazon Redshift o unas pocas llamadas API.

Conclusión

Se observa un cambio estratégico en el almacenamiento de datos debido a que las compañías migran sus soluciones y bases de datos de análisis desde soluciones locales hasta la cloud, a fin de beneficiarse de la sencillez, el desempeño y la rentabilidad que ofrece la cloud. Por medio de este documento técnico, se ofrece un informe completo del estado actual del almacenamiento de datos en AWS. Mediante AWS, se proporciona un conjunto de servicios amplio y un ecosistema

de socios fuerte que le permiten construir y ejecutar fácilmente el almacenamiento de datos empresariales en la cloud. El resultado es una arquitectura de análisis con rentabilidad y desempeño altos, capaz de escalar con su negocio, apoyándose sobre la infraestructura global de AWS.

Colaboradores

En este documento han participado las siguientes personas y organizaciones:

- Babu Elumalai, arquitecto de soluciones, Amazon Web Services
- Greg Khairallah, BDM principal, Amazon Web Services
- Pavan Pothukuchi, administrador principal de productos, Amazon Web Services
- Jim Gutenkauf, redactor senior técnico, Amazon Web Services
- Melanie Henry, editora sénior técnica, Amazon Web Services
- Chander Matrubhutam, encargado del marketing de productos, Amazon Web Services

Documentación adicional

Para obtener más ayuda, consulte las siguientes fuentes:

- Biblioteca de software Apache Hadoop[18]
- Prácticas recomendadas de Amazon Redshift[19]
- Arquitectura de Lambda[20]

Notas

1https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Data+Warehouse+Q4+2015/-/E-RES124041

2 http://aws.amazon.com/streaming-data/

3 http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html

4 http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html

5 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes

6 http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html

7 http://aws.amazon.com/redshift/partners/

8 https://aws.amazon.com/vpc/

9 https://aws.amazon.com/cloudtrail/

10 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html

11 https://aws.amazon.com/kms/

12 http://aws.amazon.com/s3/pricing/

13 http://aws.amazon.com/redshift/pricing/

14 http://docs.aws.amazon.com/redshift/latest/dg/json-functions.html

15 https://aws.amazon.com/dms/

16 https://aws.amazon.com/redshift/partners/

17 http://aws.amazon.com/ec2/spot/

18 https://hadoop.apache.org/

19 http://docs.aws.amazon.com/redshift/latest/dg/best-practices.html

20 https://en.wikipedia.org/wiki/Lambda_architecture