01.Amazon Redshift

terça-feira, 12 de setembro de 2023 13:33

Como funciona o Amazon Redshift?

O Amazon Redshift é um datawarehouse em colunas, em nuvem e totalmente gerenciado, que pode ser usado para fazer consultas analíticas e complexas e em grande conjunto de dados por meio da tecnologia de processamento massivamente paralelo (MPP).

Com suporte para ilimitado número de usuários e com o poder de ingestão de dados e análises avulsas além de relatório de painéis. Monitorando cargas e trabalho e usando Machine Learning para aprimoramento de layot físico de dados para otimizar a velocidade das consultas.

Quais os problemas que o Amazon Redshift soluciona?

O amazon redshift coleta, armazena e analisa dados estruturados ou semi-estruturados. É possível ver tendências ou análises preditivas em conjunto de dados mesmo que sejam de terceiros e você tenha permissão. Não precisa instalar, atualizar ou atualização de software já que o serviço é totalmente gerenciado.

Quais os benefícios do Amazon Redshift?

Gerenciamento de carga de trabalho - Ajuda a gerenciar com flexibilidade as prioridades de cargas de trabalho para que consultas rápidas não sejam executadas posteriormente as consultas mais longas.

Editor de consultas V2 - Usa o SQL para tornar os data lakes e dados mais acessíveis para analistas, engenheiro de dados e outros usuários de SQL com um Workbench baseado na web para exploração e análise de dados.

Design de tabela automatizado - Monitora cargas de trabalho e encontra as melhores formas de melhorar o layout físico para otimizar a velocidade das consultas.

Consulta com suas próprias ferramentas - Aumenta a flexibilidade para executar consultas no console ou conectando as ferramentas do cliente SQL como Quicksight, Tableau, Power Bi, Querybook e Jupyter notebook.

Simples interação com API - Fornece acesso a dados com diferentes tipos de aplicações tradicionais, nativos de nuvem, conteinerizados serveless e baseados em serviços web e aplicações orientadas por eventos.

Tolerante a falhas - Consegue monitorar o cluster de data warehouse e replicação dos dados caso tenha falha, além de substituir nó se necessário.

Serviços do Amazon Redshift

1. Amazon Relational Database Service (Amazon RDS) para PostgreSQL.
2. Amazon RDS para MySQL
3. Edição compatível com PostgreSQL do Amazon Aurora
4. Edição compatível com MySQL do Amazon Aurora

Com o **Amazon Redshift Spectrum** é possível consultar e recuperar com eficiência dados estruturados ou semi-estruturados de arquivos do S3, sem precisar carregar os dados em tabelas do Redshift, **O AWS Data Exchange** assina e encontra dados de terceiros, Além de tudo é possível com o Amazon Redshift ML criar, treinar e aplicar modelos de ML com SQL padrão.

Quanto custa o Amazon Redshift?

É possível escolher a configuração de cluster e tipo de nó (RA3 ou DC2), podendo optar pelos **Preços sob demanda** - Usar recursos e retomar para suspender a cobrança sob demanda quando um cluster não estiver sendo usado.
Instâncias reservadas - Estado de trabalho fixo e obter descontos significativos em relação a preços sob demanda.

Amazon Redshift Spectrum - É possível executar consultas diretamente no seu data lake no S3 de até Exabytes, pagando apenas pelos bytes verificados.

Scaling de simultaneidade - Cada usuário recebe 1 hora de créditos gratuitos de simultaneidade que é suficiente para 97% dos clientes podendo pagar apenas um valor sob demanda por segundo que exceder o crédito.

Cluster RA3 - é possível dimensionar a substituição e o armazenamento de forma independente e você só paga pelos dados que armazenam em um cluster RA3, independentemente do número de nós de computação provisionados, paga por hora pela quantidade de dados armazenados.

Amazon Redshift ML - é possível usar o SageMaker do Amazon Redshift ML para treinar os modelos.

Usando o Amazon Redshift para arquitetar uma solução de nuvem

Os principais recursos para criar um data warehouse moderno são:
- Redshift Spectrum

- Consultas federadas
- API de dados do Amazon Redshift
- Compartilhamento de dados
- Amazon Redshift ML

Passo 1

Fonte de dados on-premises
Os dados podem vir de diversas fontes como Banco de dados relacionais, Sistema de planejamento de recursos empresariais (ERP) e aplicativos web.

Fonte de dados da AWS
Os dados podem ser ingeridos a partir de serviços da AWS como o Amazon RDS, Aurora e Amazon S3.

Passo 2

Pode-se usar o comando COPY para ingerir dados para o Amazon Redshift e logo após escolher entre o **AWS Glue** ou o **Amazon Kinesis firehorse** para fazer o ETL destes dados.

Passo 3

Usar o S3 para criação de um Data Lake.

Passo 4

As consultas federadas são usadas para consultar dados dinâmicos em banco de dados externos diretamente do Amazon Redshift. As consultas federadas podem funcionar com banco de dados externos no Amazon RDS para PostgreSQL, Edição compatível com PostgreSQL do Aurora, Amazon RDS com MySQL e Edição compatível com MySQL do Aurora. Podemos usar consultas federadas para incorporar dados dinâmicos de seus banco de dados de origem como parte de seus aplicativos de BI e relatórios, pode ser feito com necessidade de carregar dados no Data Warehouse.

Passo 5

O Amazon Redshift Spectrum executa uma consulta e recupera os dados em paralelismo ao Serviço S3 da Amazon, com alta demanda de dados e velocidade elevada, grande parte do processamento dos dados é feito na camada do Amazon Redshift Spectrum e a maioria dos dados continuam no Amazon S3.

O compartilhamento dos dados pode ser feito com segurança entre uma mesma conta ou outras contas da AWS, graças ao uso do Amazon Redshift.

Passo 6

Tabela do Amazon Redshift - Com essa API é possível acessar dados do Amazon Redshift com aplicativos baseados em serviços da web incluindo o AWS Lambda e notebooks SageMaker, sem a necessidade de instalar drivers JDBC e ODBC.

A API oferece um endpoint HTTP seguro e integração com kits de desenvolvimento de software(SDK) da AWS, é possível usar o endpoint para executar instruções SQL sem gerenciar conexões.

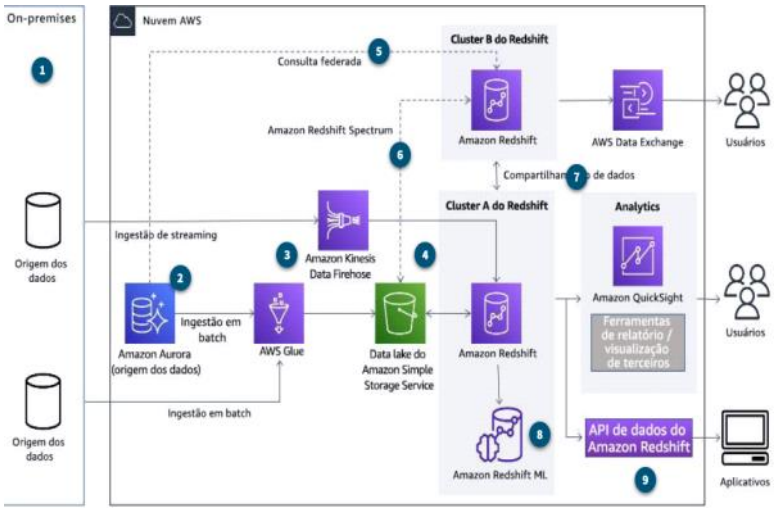

Diagrama da arquitetura do Redshift

Workload do Amazon Redshit

OLTP - é um sistema de processamento de transação online onde existem inserções, atualizações e deleções.
OLAP - Usado para principalmente para workloads analíticos, caracterizado por grandes inserções em massa, o Amazon Redshift se enquadra em OLAP.

Distribuição de dados do Amazon Redshift

Quando você cria uma tabela, pode determinar um de quatro estilos de distribuição; AUTO, EVEN, KEY ou ALL.

Se você não especificar um estilo de distribuição, o Amazon Redshift usa distribuição AUTO.

Distribuição EVEN
O nó de liderança distribui as linhas ao longo das fatias de modo round-robin, independente dos valores de qualquer coluna específica. A distribuição EVEN é apropriada quando uma tabela não participa de junções. Também é apropriado quando não há uma opção clara entre a distribuição KEY e ALL.

Distribuição KEY
As linhas são distribuídas de acordo com os valores em uma coluna. O nó líder coloca os valores correspondentes na mesma fatia do nó. Se você distribuir um par de tabelas nas chaves de união, o nó líder coloca as linhas nas fatias de acordo com os valores nas colunas de união. Desta forma, os valores correspondentes das colunas comuns são armazenados fisicamente juntos.

Distribuição ALL
Uma cópia de toda a tabela é distribuída para cada nó. Onde a distribuição EVEN ou a distribuição KEY coloca apenas uma porção das linhas da tabela em cada nó, a distribuição ALL garante que todas as linhas sejam dispostas para cada junção na qual a tabela participa.
A distribuição ALL multiplica o armazenamento exigido pelo número de nós no cluster e, portanto, ela demora muito mais tempo para carregar, atualizar ou inserir dados em várias tabelas. A distribuição ALL é apropriada somente para tabelas relativamente lentas; ou seja, tabelas que não são atualizadas frequentemente ou extensivamente. Como o custo de redistribuir tabelas pequenas durante uma consulta é baixo, não há um benefício significativo em definir tabelas de pequena dimensão como DISTSTYLE ALL.