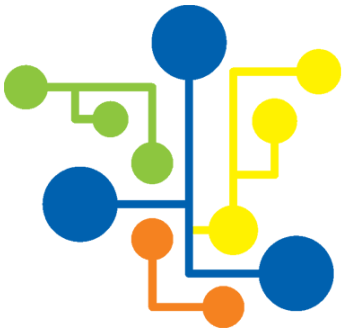


Data Science Academy

www.datascienceacademy.com.br

Design e Implementação de Data Warehouses

Arquitetura do Data Warehouse com Amazon Redshift

Configurar um Data Warehouse requer duas grandes etapas de trabalho:

1- Modelagem
2- Infraestrutura

Com relação à modelagem, os procedimentos são os mesmos independentemente da implementação on-premise ou em nuvem. Todas as etapas como definição do problema de negócio, business case, especificação, modelagem lógica, dimensional e física, seguem os mesmos padrões uma vez que estamos tratando do design do DW.

Já com relação à infraestrutura, as diferenças entre implementação on-premise e em nuvem são grandes, pois se trata de "mundos" diferentes com abordagens e custos diferentes. A escolha por uma opção ou outra, depende de diversos fatores e uma escolha profissional requer uma avaliação 360 graus.

Em uma implementação de DW em nuvem, a arquitetura envolve fatores além do serviço em nuvem sendo utilizado (Amazon Redshift em nosso caso), pois precisamos definir níveis de acesso e segurança, armazenamento, backup, integração, ETL, custos de transferência de dados, largura de banda para acesso aos serviços em nuvem, etc... Tudo isso deve ser levado em consideração para que o custo de uma implementação em nuvem justifique sua utilização.

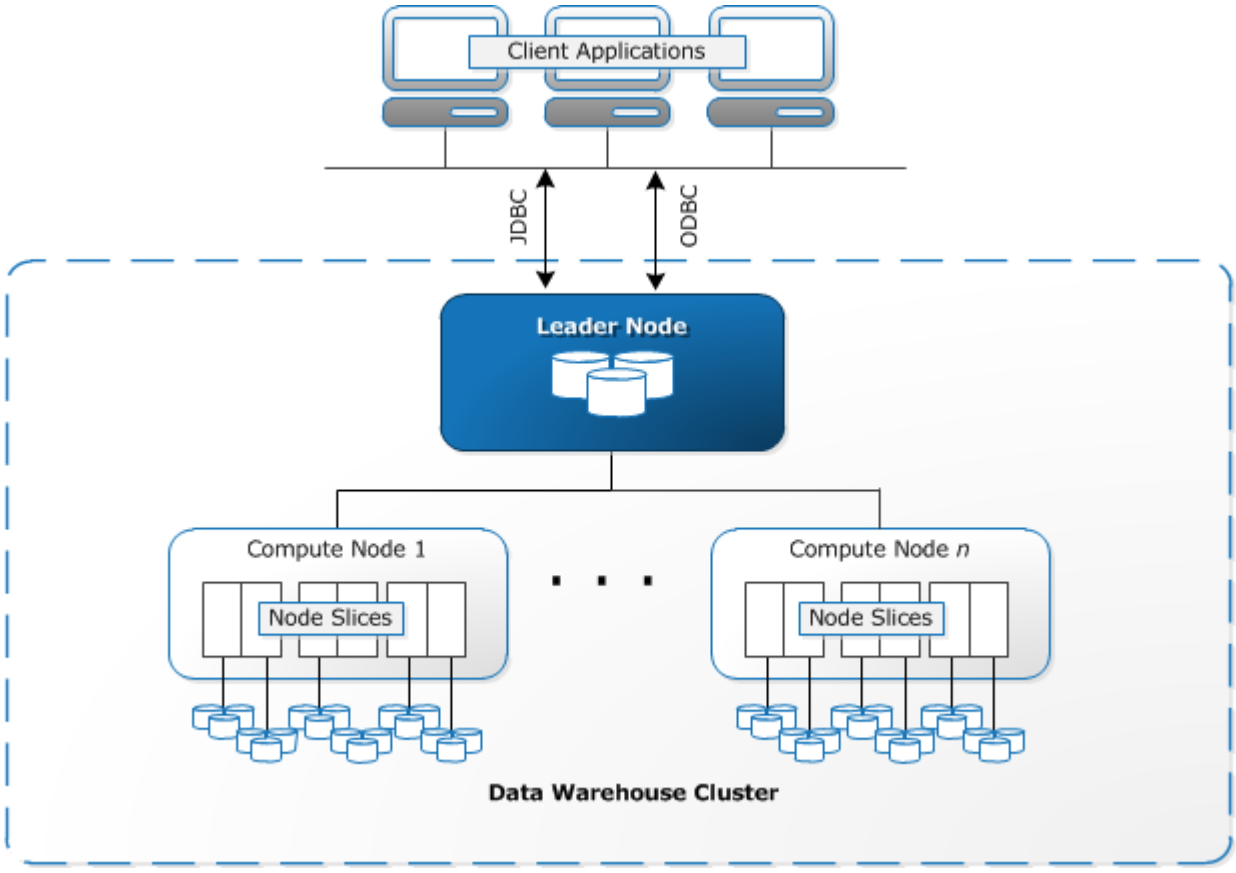
Embora existam outras opções, podemos falar especificamente da AWS, uma vez que estamos usando este serviço aqui em nosso projeto. Além do Amazon Redshift, uma implementação de DW em nuvem (normalmente) requer o uso de outros serviços AWS, tais como:

- S3 (Simple Storage Service)
- VPC (Virtual Private Cloud)
- EIP (Elastic IP)
- Cloud Watch
- IAM (Identity and Access Management)

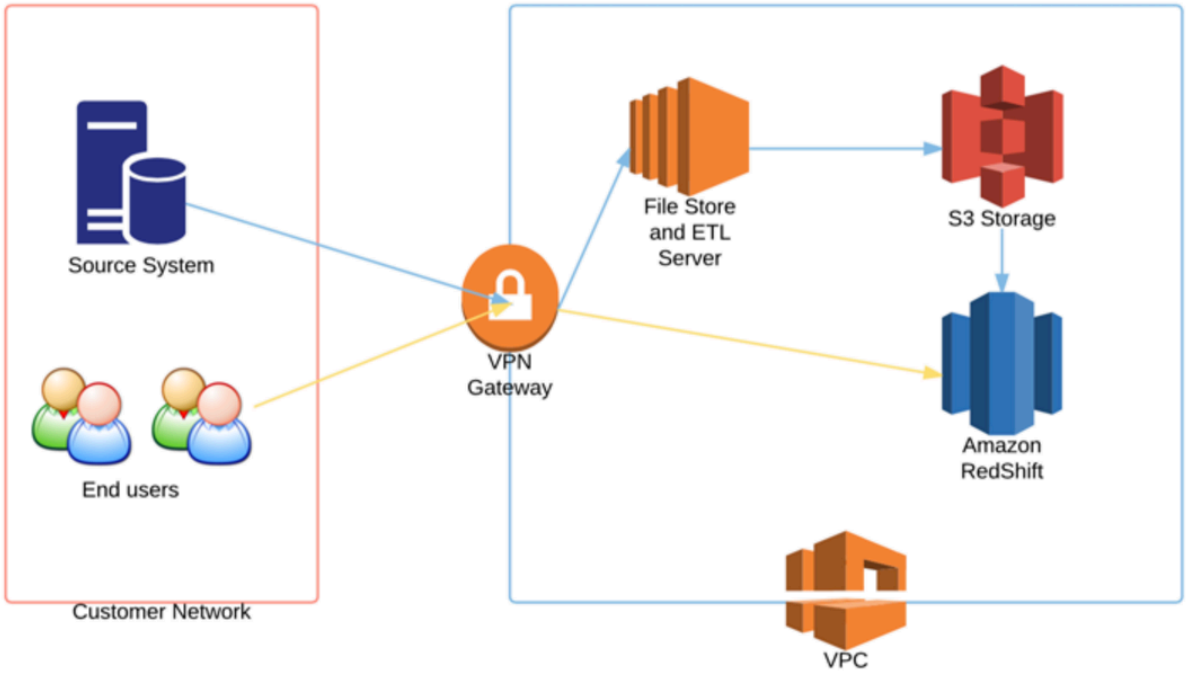
Destes 5 serviços, apenas o IAM não possui custo. Embora seja possível usar uma Default VPC (como você verá durante as próximas aulas), qualquer customização poderá gerar cobranças adicionais. Armazenamento com o S3 (para

colocar os arquivos que serão carregados no DW), EIP (para ter um IP exclusivo) e Cloud Watch (para monitoramento e alarmes) não são obrigatórios em uma implementação do Amazon Redshift, mas podem ajudar colaboram para tornar a experiência em nuvem bem mais profissional.

A arquitetura do DW na AWS, pode ainda ser dividida em duas etapas. Primeiro a arquitetura do Cluster Amazon Redshift (conforme imagem abaixo). O número de compute nodes determina o custo do cluster e deve ser avaliado com cuidado.



Em segundo, a arquitetura de acesso ao Cluster Amazon Redshift, para que os usuários tenham acesso ao DW, conforme imagem abaixo:



Neste momento, outros serviços AWS devem ser considerados. O VPN Gateway permite criar uma conexão segura entre o escritório da empresa e o ambiente AWS, mas também é um serviço AWS com custo à parte do Redshift.

É função do Engenheiro de Dados pensar na melhor arquitetura para o projeto, visando reduzir custos e otimizar a utilização de recursos. Embora seja possível usar o Redshift apenas por um período de tempo e pagar pelas horas de uso, nem todos os serviços seguem este mesmo padrão (não faz sentido gravar e apagar dados no S3 por exemplo, sendo o ideal usar suas funcionalidades de versionamento). Arquiteturas mais avançadas podem requerer o uso de serviços de mensageria, como o AWS SQS, que garante o funcionamento da aplicação mesmo em caso de queda do DW. O SQS (como você deve imaginar) também possui custo adicional.

Recomendamos o uso da calculadora AWS (mostrada no capítulo anterior) para auxiliar na precificação da arquitetura DW em nuvem com AWS.

Nos próximos vídeos veremos parte desta arquitetura ao implementarmos o DW com o Amazon Redahift.