

VINÍCIUS HENRIQUE DOS SANTOS

POSTECH

MACHINE LEARNING ENGINEERING

BIG DATA STORAGE STRUCTURES

AULA 01

SUMÁRIO

| | |
|----------------------------------|----|
| O QUE VEM POR AÍ? | 3 |
| HANDS ON | 4 |
| SAIBA MAIS..... | 5 |
| O QUE VOCÊ VIU NESTA AULA? | 28 |
| REFERÊNCIAS..... | 29 |

EMAN

O QUE VEM POR AÍ?

Como você faria para analisar uma centena de linhas em um banco de dados? E para analisar milhares? Um tanto quanto desafiador, não? Prepare-se... você está prestes a entender o funcionamento e a arquitetura por trás de um dos sistemas mais robustos para bancos de dados voltados à análise, o Amazon Redshift.

Você vai entender os conceitos de um sistema OLAP e as diferenças para um OLTP, bem como aprenderá a configurar os serviços do Redshift para te ajudarem com qualquer tarefa analítica que possa vir a ter! Aperte os cintos, pois nessa aula você trabalhará com um volume de dados na casa de centenas de milhões de linhas de uma forma muito rápida, eficiente e performática!

HANDS ON

Nesse Hands On você terá uma tarefa que te levará do zero até conseguir manipular volumes de dados na grandeza de centenas de milhões de linhas dentro do Amazon Redshift.

1. Realizar as configurações de conta:
 - a. Vpc e subnet.
 - b. Criação de roles.
 - c. Security Group.
 - d. Criação de buckets S3.
2. Configurar um cluster do Amazon Redshift.
3. Instanciar o cluster.
4. Realizar a ingestão de dados de fonte externa (S3).
5. Realizar consultas analíticas sobre os dados.

SAIBA MAIS

Modelagem de dados OLAP

Em um repositório de dados voltado para análises, também conhecido como OLAP, nós temos três componentes principais que permitem transformar a modelagem feita para transações em uma modelagem voltada para análises. São eles: data warehouses, tabelas fato e tabelas dimensão.

Data warehouses

Um data warehouse é o componente central de qualquer sistema voltado para análises, visto que tem como finalidade armazenar os dados de diversas fontes diferentes de forma organizada para consultas. Diferentemente dos bancos de dados operacionais, que são otimizados para transações eficientes e atualização de registros (OLTP), os data warehouses são projetados para consultas e análises rápidas em grandes volumes de dados, em que o objetivo é trazer respostas importantes para o negócio.

Os dados armazenados no DW são extraídos de diversas fontes internas ou externas da empresa, passando por um processo de organização, limpeza, ganho de significado, garantia de consistência e cálculos de valores. Esse processo é conhecido por ETL e roda dentro do DW realizando a extração dos dados nas origens, a transformação e a carga (Extract, Transform and Load, no inglês).

Os dados internos da própria empresa vêm de sistemas como CRM, cadastro, faturamento, logística, telecom e muitos outros; além dos internos existem muitas possibilidades de aquisição de dados externos, como redes sociais, leads de marketing, score de mercado como bureaus de crédito e muitos outros.

Os data warehouses frequentemente utilizam o modelo “dimensional” para organizar os dados em fatos e dimensões.

Tabelas fato e dimensões

O modelo estrela é um dos designs mais comuns e simplificados para esquemas de data warehouses. Sua estrutura é projetada para otimizar consultas rápidas e eficientes, tornando-o particularmente adequado para aplicações de BI e análise de dados.

Componentes principais:

- **Tabela Fato:** no centro do modelo estrela, a tabela fato armazena dados quantitativos relacionados a eventos ou transações de negócios, como vendas, custos e lucros. Esses dados são tipicamente numéricos e podem ser agregados para análise (por exemplo, somar vendas totais). A tabela fato contém chaves estrangeiras que apontam para várias tabelas de dimensões relacionadas.
- **Tabelas de Dimensão:** cercando a tabela fato, as tabelas de dimensão contêm dados contextuais sobre as métricas na tabela fato. Estas tabelas armazenam atributos descritivos, ou dimensões, que caracterizam os fatos. Por exemplo: uma tabela de dimensão "Produto" pode incluir informações como nome do produto, categoria e preço. Outras tabelas de dimensão comuns incluem Cliente, Tempo, e Localização.

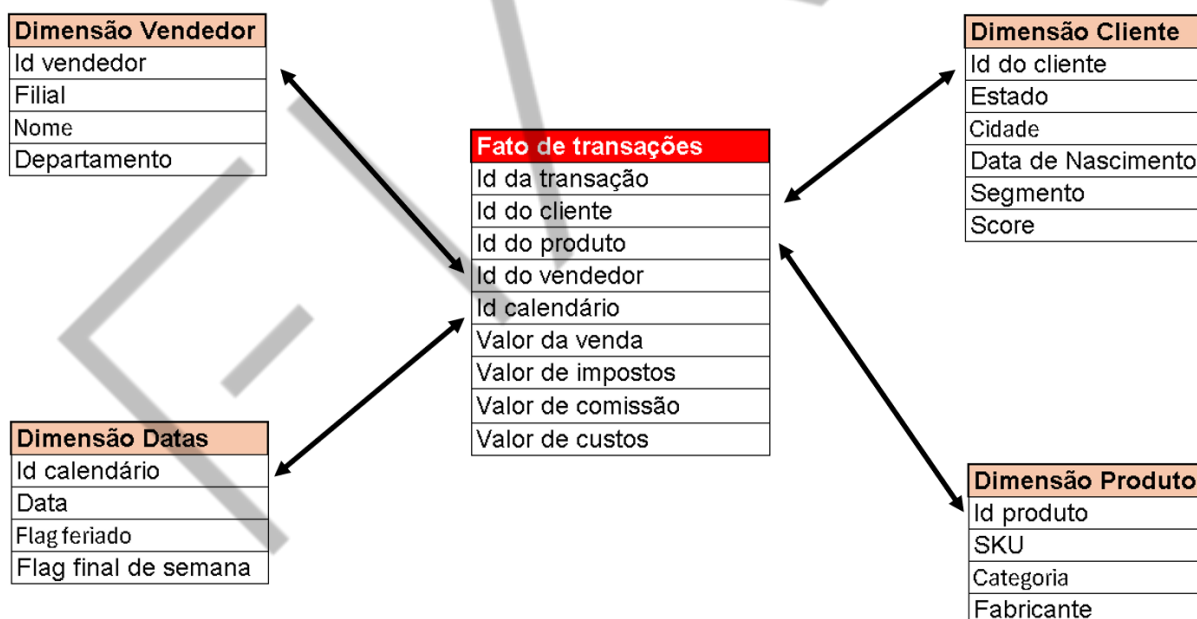


Figura 1 – Tabela de Dimensão
Fonte: Elaborado pelo autor (2024)

Benefícios do AWS Redshift

1) Foco na geração de valor: a proposta principal dessa ferramenta é permitir que as empresas consigam investir energia em geração de valor através dos dados e não gastem energia com as demandas de infraestrutura. Sendo uma ferramenta totalmente gerenciada pela cloud AWS, questões como aplicação de paths ou atualização de softwares e conectores não são preocupação de pessoas engenheiras, analistas e desenvolvedoras.

2) Análise de TODOS os seus dados: com o Amazon Redshift, você pode executar analytics em dados complexos e dimensionáveis em seus bancos de dados operacionais, data lakes, data warehouses e milhares de conjuntos de dados de terceiros.

Com sua nova capacidade de consulta federada, você pode acessar seu banco de dados relacional operacional e consultar dados dinâmicos em um ou mais bancos de dados nos seguintes serviços sem a necessidade de movimentar dados:

- a. Amazon Relational Database Service (Amazon RDS) para PostgreSQL.
- b. Amazon RDS para MySQL.
- c. Edição compatível com PostgreSQL do Amazon Aurora.
- d. Edição compatível com MySQL do Amazon Aurora.

Com o Amazon Redshift Spectrum, você pode consultar e recuperar com eficiência dados estruturados e semiestruturados de arquivos no Amazon Simple Storage Service (Amazon S3).

Não é preciso carregar os dados em tabelas do Amazon Redshift. Com o AWS Data Exchange para o Amazon Redshift, é possível encontrar e assinar dados de terceiros no AWS Data Exchange, os quais podem ser consultados rapidamente em um data warehouse do Amazon Redshift.

Você pode usar o Amazon Redshift ML para criar, treinar e aplicar modelos de ML com SQL padrão.

3) Conectividade: o Amazon Redshift oferece flexibilidade para você executar consultas diretamente no console da solução. Isso é feito usando o editor de consultas v2, que é um workbench baseado na web para exploração, análise e criação de

gráficos de dados. O Amazon Redshift também fornece drivers de conectividade de banco de dados Java (JDBC) e conectividade de banco de dados aberto (ODBC) para se conectar com ferramentas de cliente SQL (Power BI, Tableau, Jupyter Notebook, ferramentas de ETL).

4) Suporte a formatos de arquivos diversos: o Amazon Redshift tem a capacidade de processar e consultar arquivos em diversos formatos, entre eles:

- Parquet;
- ORC;
- JSON;
- Avro;
- CSV.

5) Desempenho: o Amazon Redshift oferece desempenho para consultas rápidas em conjuntos de dados que vão de gigabytes a petabytes. Armazenamento colunar, compactação de dados e mapas de zonas são elementos que reduzem a quantidade necessária de E/S para executar as consultas.

Além de codificações padrão na indústria, como LZO e Zstandard, o Amazon Redshift oferece codificação de compactação para fins específicos, AZ64, para tipos numéricos e de data/horário. Essa codificação foi desenvolvida para oferecer economia de armazenamento e desempenho de consulta otimizado.

O Amazon Redshift oferece desempenho rápido de forma constante, mesmo com milhares de consultas simultâneas. Você pode consultar dados no data warehouse do Amazon Redshift ou diretamente no data lake do Amazon S3. Além disso, o scaling de simultaneidade do Amazon Redshift suporta um número praticamente ilimitado de consultas e usuários simultâneos, mantendo níveis constantes de serviço e adicionando capacidade transitória em questão de segundos à medida que a simultaneidade aumenta.

O Amazon Redshift usa cache de resultado para entregar tempos de resposta de milissegundos para consultas repetidas. Assim, painel, visualização e ferramentas de BI que executam consultas repetidas podem ter um ganho significativo de desempenho.

6) Segurança: a AWS tem funcionalidades abrangentes de segurança para atender aos mais exigentes requisitos e o Amazon Redshift fornece segurança de dados pronta para uso sem custo adicional. Com as configurações de parâmetros, o Amazon Redshift pode ser configurado para usar Secure Sockets Layer (SSL) para proteger dados em trânsito e criptografia Advanced Encryption Standard (AES)-256 para dados em repouso. Você pode configurar as definições do firewall e controlar o acesso da rede ao seu cluster de data Warehouse.

É possível executar o Amazon Redshift dentro do Amazon Virtual Private Cloud (Amazon VPC) para isolar o cluster de data warehouse na sua própria rede virtual. Você pode conectá-la à sua infraestrutura de TI existente por meio de uma rede privada virtual (VPN) IPsec criptografada padrão do setor.

O Amazon Redshift também se integra com o AWS CloudTrail. Portanto, todas as chamadas da interface de programação do aplicativo (API) do Amazon Redshift podem ser auditadas. O Amazon Redshift registra todas as operações SQL, incluindo tentativas de conexão, consultas e alterações feitas no seu data warehouse. Você pode acessar esses logs usando consultas SQL em relação a tabelas do sistema ou salvar os logs em um local seguro no Amazon S3.

Arquitetura e casos de uso para o AWS redshift

O Amazon Redshift usa SQL para analisar dados estruturados e semiestruturados em data warehouses, bancos de dados operacionais e data lakes. O hardware desenvolvido pela AWS alinhado com ferramentas de Machine Learning dentro dessa solução ajudam a entregar um melhor desempenho em termos de custos para as organizações e, por conta disso, ele está ganhando cada vez mais espaço e funções-chave em uma arquitetura de dados moderna.

Veja o exemplo de arquitetura em que o AWS Redshift faz o papel de data warehouse interagindo com inúmeros outros serviços e bancos de dados fora da cloud e entregando os dados de forma transparente para usuários:

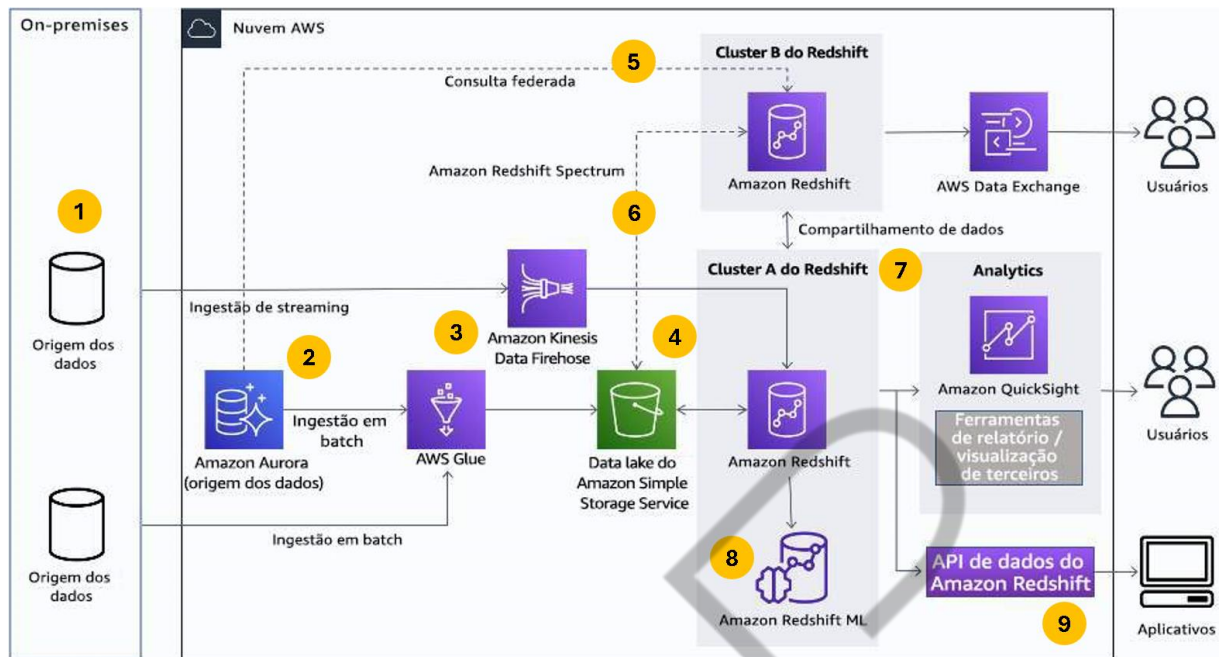


Figura 2 – Exemplo de Arquitetura cloud AWS

Fonte: [Google Imagens](#) (2024)

1. **Fontes de dados on-premises:** os dados podem ser ingeridos a partir de fontes de dados on-premises. Essas fontes podem ser bancos de dados relacionais, arquivos gerados por sistemas em algum diretório, dados disponibilizados por ERPs, aplicativos WEB etc.
2. **Fontes de dados na AWS:** os dados também podem ser ingeridos a partir de fontes AWS, como Amazon RDS, Aurora e S3
3. **Ingestão dos dados:** nessa arquitetura estamos usando dois serviços para possibilitar o carregamento dos dados de fontes externas. A estratégia pode variar de acordo com a particularidade da empresa e do negócio, podendo trazer os dados de uma forma bruta (exatamente como são na origem) ou já aplicando tratamentos, transformações e dando significado.
 - a. **AWS Glue:** nessa arquitetura o AWS Glue pode ser usado para integrar os ambientes, realizando cópia dos dados, preparação e modelos de machine learning, entre outras coisas.
 - b. **AWS Kinesis Data Firehose:** nesse cenário essa ferramenta entrega funcionalidades voltadas à ingestão de dados em tempo real, ou seja: a medida em que o sistema origem disponibiliza um novo dado, ele é

automaticamente consumido pelo Kinesis Data Firehose e ingerido para o AWS Redshift.

4. **Data lake (AWS S3):** o AWS S3 é um serviço de armazenamento de dados em grande escala e desempenho para dados estruturados e não estruturados, sendo a melhor opção para armazenamento de dados na criação de data lakes. Suas vantagens dizem respeito à capacidade ilimitada de escalabilidade em tamanho de forma eficaz em relação ao custo, em um ambiente seguro no qual os dados são protegidos com uma garantia de durabilidade em 99,9999999999% (11 noves).
5. **Consultas federadas:** esse recurso do Amazon Redshift permite que você realize consultas em ambientes externos, ou seja, outros bancos de dados diretamente de dentro do Redshift. Esse recurso funciona com Postgres e MySQL. Essa funcionalidade permite configuração de dados externos como parte das cargas de trabalho dentro do Redshift sem a necessidade de desenvolvimento e configuração de uma rotina ETL para transferência e tratamento dos dados.
6. **Amazon Redshift Spectrum:** com o Redshift Spectrum é possível configurar consultas e recuperar dados de arquivos armazenados no S3 de forma eficiente, sem a necessidade de carregar esses dados em tabelas dentro do AWS Redshift. Esse recurso realiza as consultas de forma massiva e paralela para que seja possível analisar grandes volumes de dados em alta velocidade.
7. **Compartilhamento de dados:** permite o compartilhamento de dados entre serviços, ferramentas e usuários dentro ou fora da região, bem como compartilhamento com usuários de contas distintas na AWS.
8. **Amazon Redshift ML:** especialistas em desenvolvimento de bancos de dados e analistas de dados podem usar o Redshift para criar, treinar e aplicar modelos de Machine Learning. O Redshift ML permite fazer uso do Amazon SageMaker para construção e desenvolvimento de quaisquer tipos de modelos de Machine Learning.
9. **API de dados do Amazon Redshift:** essa API permite que consultas SQL sejam realizadas por aplicativos ou serviços WEB, incluindo Lambda e

SageMaker, sem a necessidade de instalações de drivers JDBC ou ODBC. Essa API oferece um endpoint HTTP seguro e integração com Kits de Desenvolvimento de Software (SDK).

Parte da excelente performance com grande volume de dados do Amazon Redshift se explica através do armazenamento de dados em forma colunar.

O armazenamento colunar nas tabelas dos bancos de dados permite que o Redshift nos traga respostas tão rápidas em todas as consultas que realizamos porque essa técnica reduz imensamente a quantidade de leituras/gravações de dados em disco, o que frequentemente é o gargalo na performance da maioria dos sistemas.

Veja nos exemplos a seguir como essa técnica converte a eficiência gerada em economia de tempo, recursos computacionais e dinheiro.

| SSN | Name | Age | Addr | City | St |
|-----------|-------|-----|---------------|---------|----|
| 101259797 | SMITH | 88 | 899 FIRST ST | JUNO | AL |
| 892375862 | CHIN | 37 | 16137 MAIN ST | POMONA | CA |
| 318370701 | HANDU | 12 | 42 JUNE ST | CHICAGO | IL |

| | | |
|---|---|--|
| 101259797 SMITH 88 899 FIRST ST JUNO AL | 892375862 CHIN 37 16137 MAIN ST POMONA CA | 318370701 HANDU 12 42 JUNE ST CHICAGO IL |
| Block 1 | Block 2 | Block 3 |

Figura 3 – Exemplo de armazenamento tradicional
Fonte: AWS (2024)

Essa imagem nos mostra como os dados de uma tabela são armazenados em “gavetas” no disco em linhas.

Dessa forma, nos bancos de dados relacionais típicos cada linha contém valores para um único registro, ou seja: os valores são armazenados sequencialmente para compor os dados de todas as colunas do registro de forma horizontal. Entretanto, se a quantidade (ou tamanho da linha) de dados de um registro for maior do que o espaço de armazenamento de um bloco no disco, o armazenamento de um registro invariavelmente vai ocupar mais de uma linha.

Em contrapartida, se a quantidade de dados do registro for menor do que o tamanho do bloco de armazenamento do disco teremos um problema de uso ineficiente do storage de armazenamento.

Normalmente, em sistemas OLTP as atualizações de dados envolvem um registro (ou uma pequena quantidade) por vez e as atualizações são aplicadas para várias colunas do mesmo registro; por conta disso entende-se que o armazenamento em linhas funciona para sistemas orientados a transação e não a análises.

A figura 4 ilustra o armazenamento orientado a colunas nos blocos do disco, em que os dados das colunas são armazenados sequencialmente.

| SSN | Name | Age | Addr | City | St |
|-----------|-------|-----|---------------|---------|----|
| 101259797 | SMITH | 88 | 899 FIRST ST | JUNO | AL |
| 892375862 | CHIN | 37 | 16137 MAIN ST | POMONA | CA |
| 318370701 | HANDU | 12 | 42 JUNE ST | CHICAGO | IL |

| | | | | | | | | | | | | | | | | | | | | | | |
|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|
| 101259797 | | 892375862 | | 318370701 | | 468248180 | | 378568310 | | 231346875 | | 317346551 | | 770336528 | | 277332171 | | 455124598 | | 735885647 | | 387586301 |
|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|--|-----------|

Block 1

Figura 4 –Exemplo de armazenamento colunar
Fonte: AWS (2024)

Ao receber cada registro, o Amazon Redshift conforme de forma transparente os dados em armazenamento colunar. Neste modelo simplificado, quando se adota o armazenamento por colunas, cada conjunto de dados armazena valores de uma mesma coluna até três vezes mais eficientemente que o método tradicional baseado em linhas.

Isso implica que acessar a mesma quantidade de valores de coluna para um número idêntico de registros demanda apenas um terço das operações de leitura/escrita se comparado ao método de armazenamento por linhas. Em cenários reais, especialmente em tabelas com uma grande quantidade de colunas e registros, a economia de armazenamento se torna ainda mais significativa.

Outro benefício é que, visto que cada conjunto contém dados de um único tipo, é possível aplicar métodos de compressão específicos para esses dados, o que

resulta em uma redução adicional tanto do uso de espaço em disco quanto das operações de leitura/escrita.

Criação de role, VPC E Security Group Redshift

Abra sua console da AWS e procure o serviço IAM:

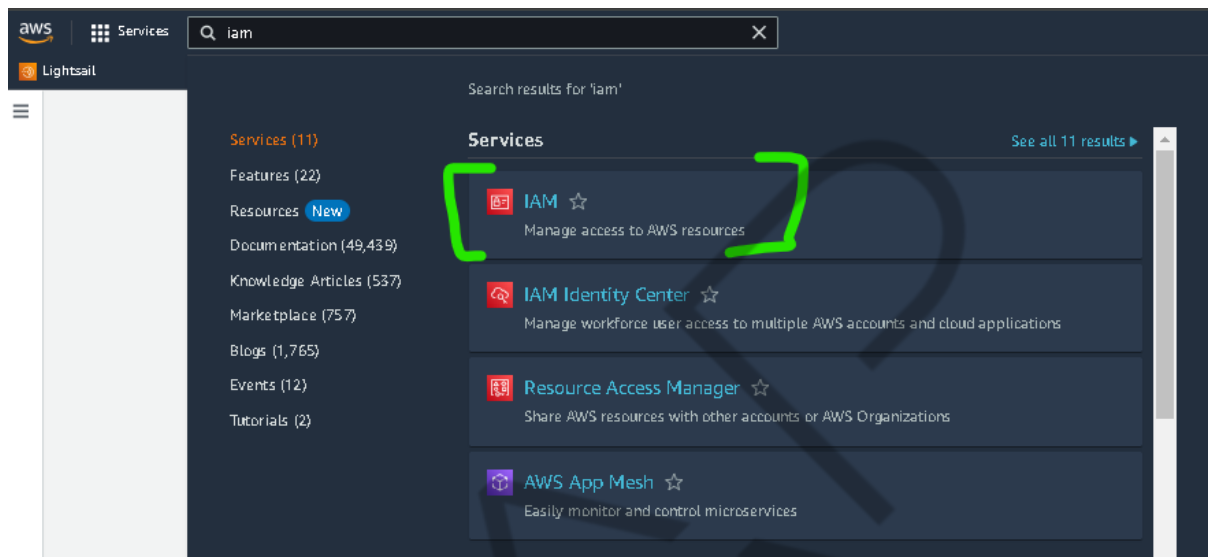


Figura 5 – Serviço IAM
Fonte: Fonte: Elaborado pelo autor (2024)

Clique em “Roles”:

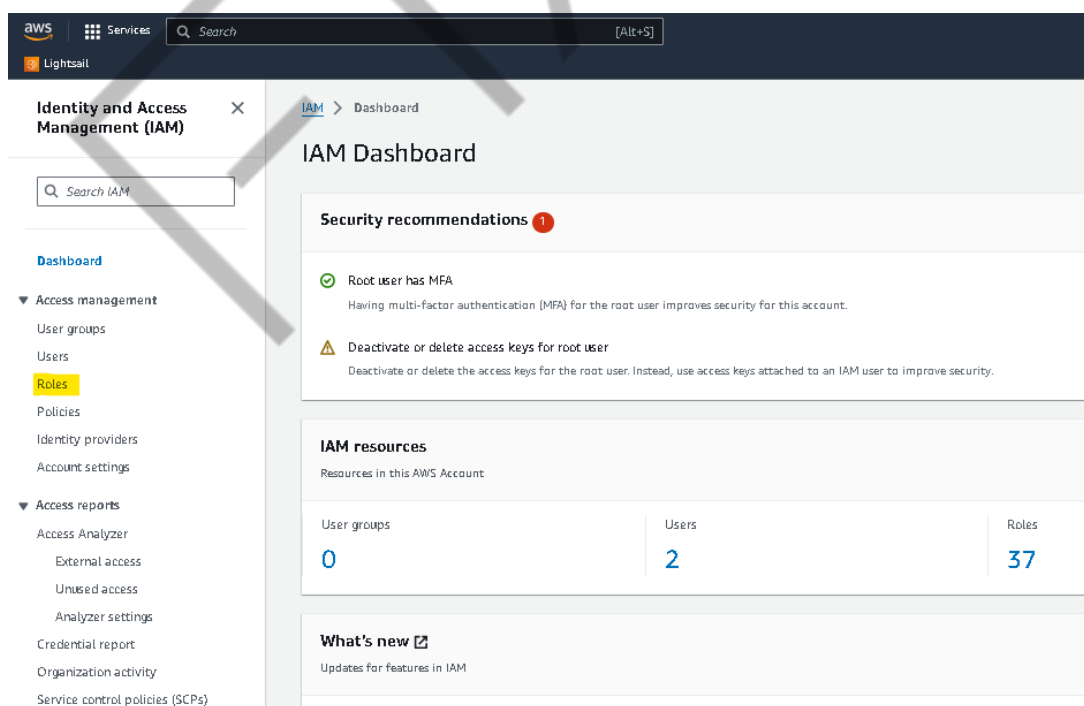


Figura 6 – Roles
Fonte: Elaborado pelo autor (2024)

Clique em “Create Role” no canto superior direito:

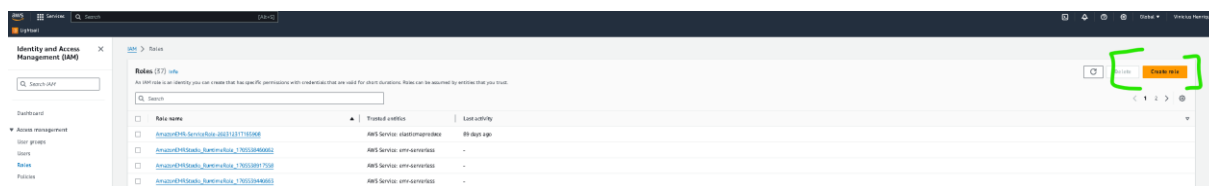


Figura 7 – Create Role (1)
Fonte: Elaborado pelo autor (2024)

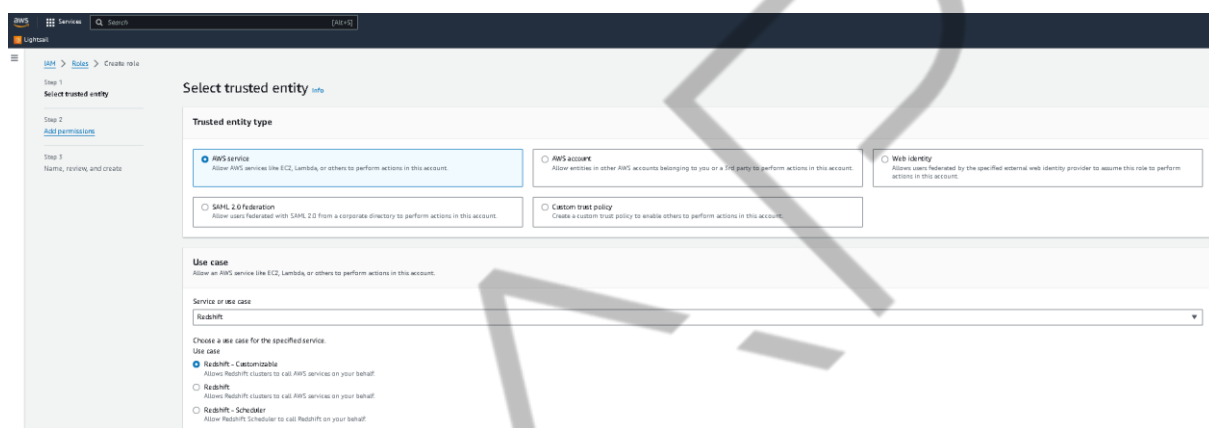


Figura 8 – Create Role (2)
Fonte: Elaborado pelo autor (2024)

Selecione as seguintes políticas:

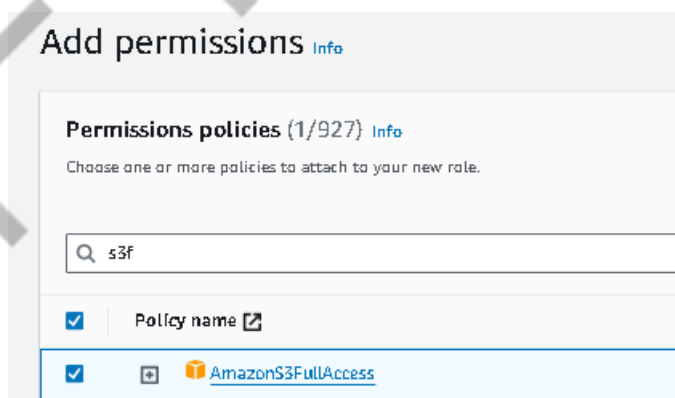


Figura 9 – Seleção de Políticas (1)
Fonte: Elaborado pelo autor (2024)

The screenshot shows the 'Add permissions' interface. At the top, it says 'Add permissions' with an 'Info' link. Below this, it says 'Permissions policies (2/927)' with an 'Info' link. A subtitle reads 'Choose one or more policies to attach to your new role.' There is a search bar containing 'ec2f'. Below the search bar, there is a table with two columns: a checkbox and 'Policy name'. The first row shows a checked checkbox and the policy name 'AmazonEC2FullAccess'.

| <input checked="" type="checkbox"/> | Policy name |
|-------------------------------------|---------------------|
| <input checked="" type="checkbox"/> | AmazonEC2FullAccess |

Figura 10 – Seleção de Políticas (2)
Fonte: Elaborado pelo autor (2024)

The screenshot shows the 'Add permissions' interface. At the top, it says 'Add permissions' with an 'Info' link. Below this, it says 'Permissions policies (3/927)' with an 'Info' link. A subtitle reads 'Choose one or more policies to attach to your new role.' There is a search bar containing 'redshiftf'. Below the search bar, there is a table with two columns: a checkbox and 'Policy name'. The first row shows a checked checkbox and the policy name 'AmazonRedshiftFullAccess'.

| <input checked="" type="checkbox"/> | Policy name |
|-------------------------------------|--------------------------|
| <input checked="" type="checkbox"/> | AmazonRedshiftFullAccess |

Figura 11 – Seleção de Políticas (3)
Fonte: Elaborado pelo autor (2024)

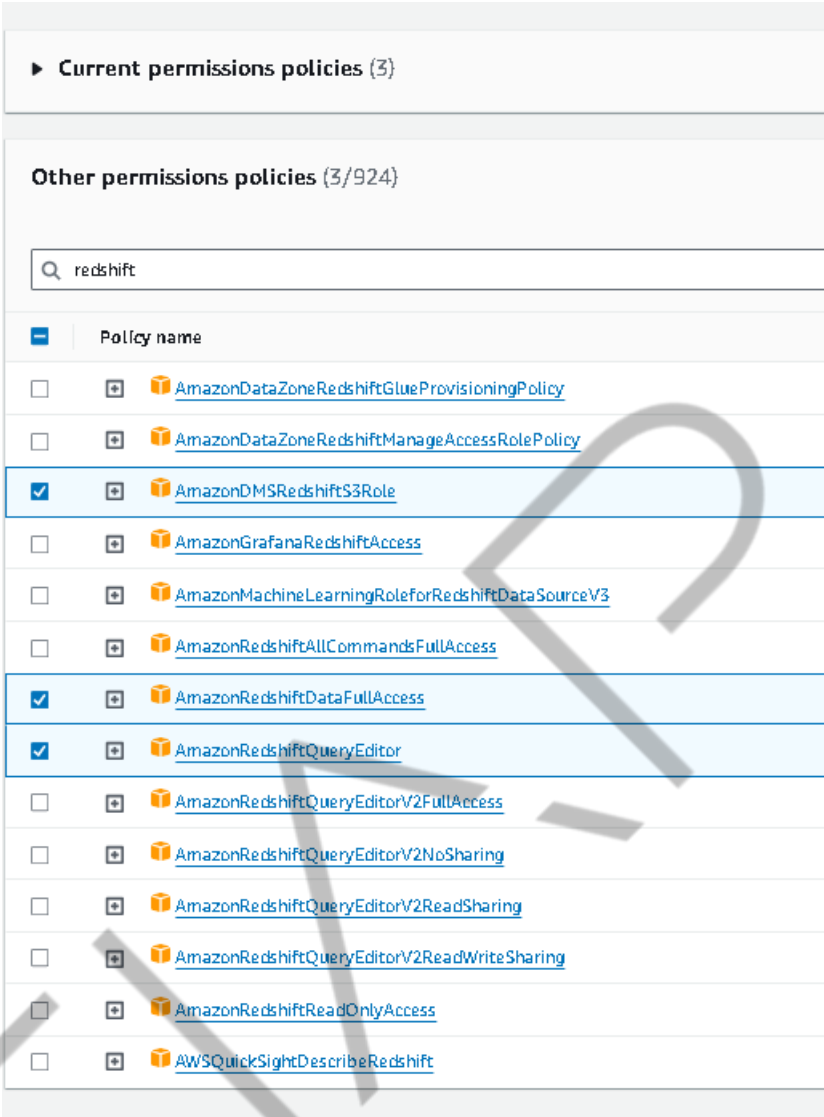


Figura 12 – Seleção de Políticas (4)
Fonte: Elaborado pelo autor (2024)

Siga para a próxima página:

Name, review, and create

Role details

Role name
Enter a meaningful name to identify this role.

Redshift-Role

Maximum 64 characters. Use alphanumeric and '+=, @, -, _' characters.

Description
Add a short explanation for this role.

Regra criada para o Lab FIAP

Maximum 1000 characters. Use alphanumeric and '+=, @, -, _' characters.

Step 1: Select trusted entities

Trust policy

```
1 {  
2   "Version": "2012-10-17",  
3   "Statement": [  
4     {  
5       "Effect": "Allow",  
6       "Action": [  
7         "sts:AssumeRole"  
8       ],  
9       "Principal": {  
10        "Service": [  
11          "redshift.amazonaws.com"  
12        ]  
13      }  
14    }  
15  ]  
16 }
```

Figura 13 – Regras e políticas de acesso
Fonte: Elaborado pelo autor (2024)

Step 2: Add permissions

Permissions policy summary

| Policy name | Type | Attached as |
|---------------------------------|-------------|--------------------|
| AmazonELBAccess | AWS managed | Permissions policy |
| AmazonEC2Access | AWS managed | Permissions policy |
| AmazonS3Access | AWS managed | Permissions policy |

Step 3: Add tags

Add tags - optional

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

No tags associated with this resource.

Add new tag

You can add up to 50 more tags.

Cancel

Previous

Create role

Figura 14 – Regras e políticas de acesso aplicadas
Fonte: Elaborado pelo autor (2024)

Clique em “create role”.

VPC

Caso você não tenha nenhuma VPC configurada, encontre o serviço de VPCs no console da AWS:

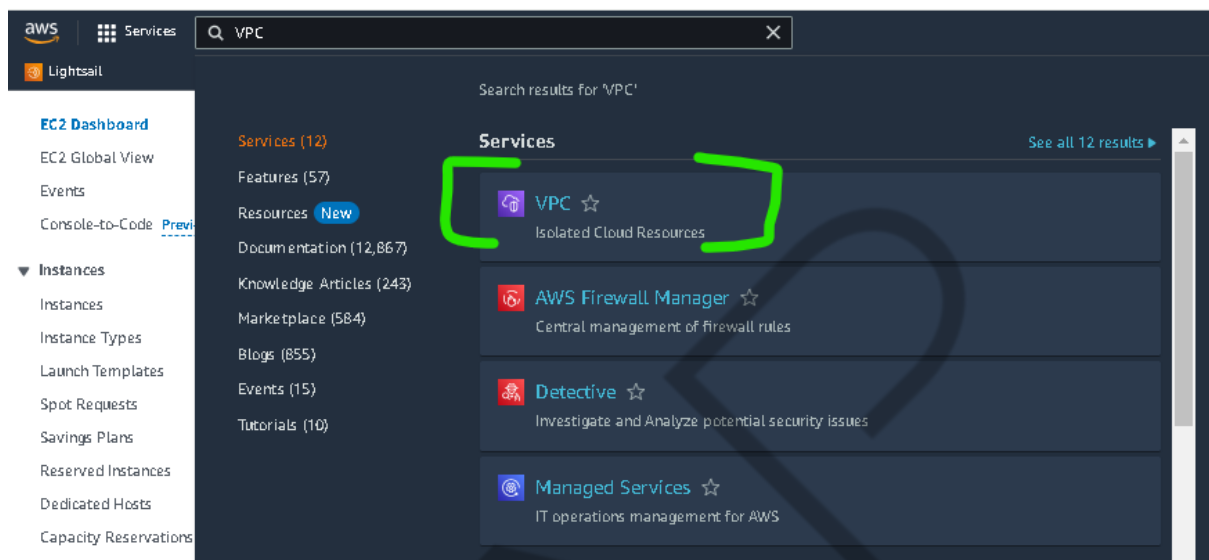


Figura 15 – VPC
Fonte: Elaborado pelo autor (2024)

Vá em “your VPCs”:

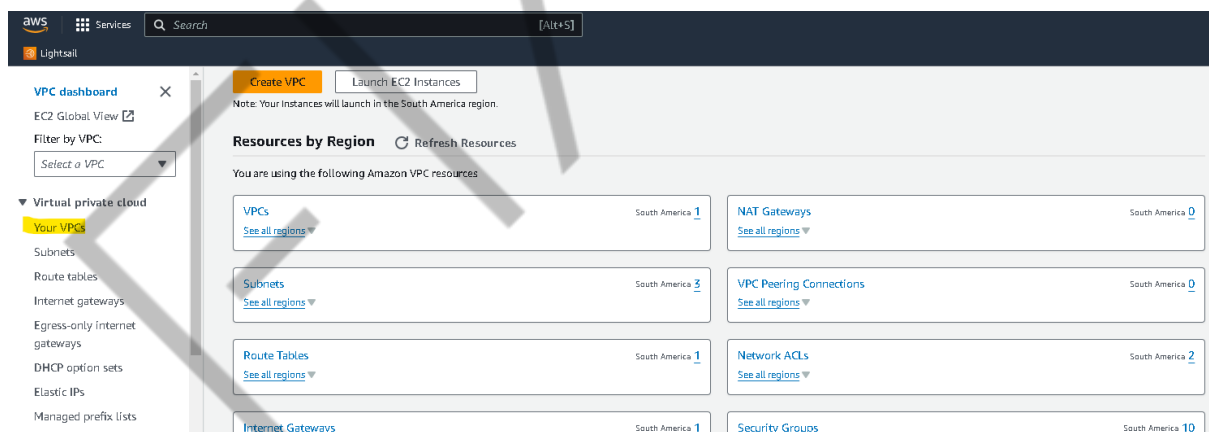


Figura 16 – Suas VPCs
Fonte: Elaborado pelo autor (2024)

Clique em “Create VPC” e siga as configurações:

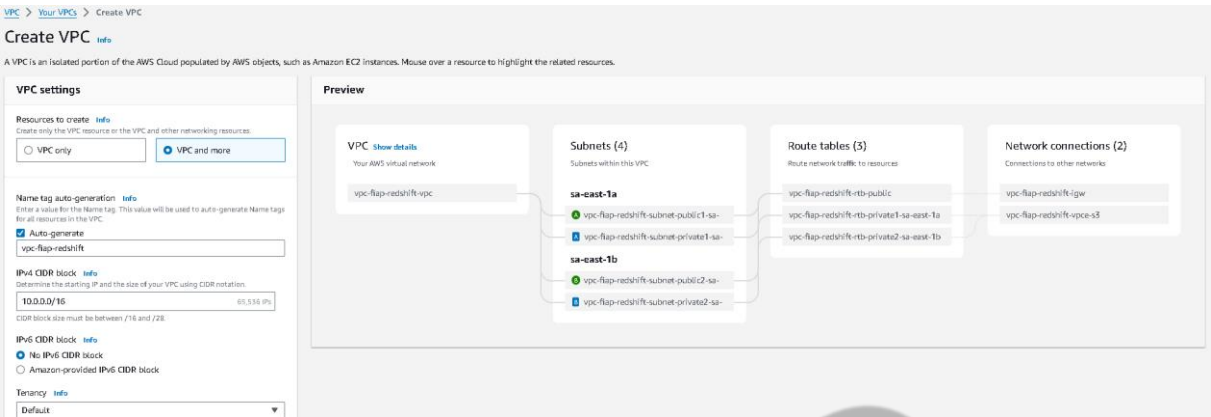


Figura 17 – Criando uma VPC (1)
Fonte: Elaborado pelo autor (2024)

Number of Availability Zones (AZs) [Info](#)
Choose the number of AZs in which to provision subnets. We recommend at least two AZs for high availability.

1 2 3

► Customize AZs

Number of public subnets [Info](#)
The number of public subnets to add to your VPC. Use public subnets for web applications that need to be publicly accessible over the internet.

0 2

Number of private subnets [Info](#)
The number of private subnets to add to your VPC. Use private subnets to secure backend resources that don't need public access.

0 2 4

► Customize subnets CIDR blocks

NAT gateways (\$) [Info](#)
Choose the number of Availability Zones (AZs) in which to create NAT gateways. Note that there is a charge for each NAT gateway.

None In 1 AZ 1 per AZ

VPC endpoints [Info](#)
Endpoints can help reduce NAT gateway charges and improve security by accessing S3 directly from the VPC. By default, full access policy is used. You can customize this policy at any time.

None S3 Gateway

DNS options [Info](#)

- ☒ Enable DNS hostnames
- ☒ Enable DNS resolution

► Additional tags

Cancel Create VPC

Figura 18 – Criando uma VPC (2)
Fonte: Elaborado pelo autor (2024)

Clique em “Create VPC”:

Number of Availability Zones (AZs) [Info](#)
Choose the number of AZs in which to provision subnets. We recommend at least two AZs for high availability.

1 2 3

► Customize AZs

Number of public subnets [Info](#)
The number of public subnets to add to your VPC. Use public subnets for web applications that need to be publicly accessible over the internet.

0 2

Number of private subnets [Info](#)
The number of private subnets to add to your VPC. Use private subnets to secure backend resources that don't need public access.

0 2 4

► Customize subnets CIDR blocks

NAT gateways (\$) [Info](#)
Choose the number of Availability Zones (AZs) in which to create NAT gateways. Note that there is a charge for each NAT gateway.

None In 1 AZ 1 per AZ

VPC endpoints [Info](#)
Endpoints can help reduce NAT gateway charges and improve security by accessing S3 directly from the VPC. By default, full access policy is used. You can customize this policy at any time.

None S3 Gateway

DNS options [Info](#)

- ☒ Enable DNS hostnames
- ☒ Enable DNS resolution

► Additional tags

Cancel Create VPC

Figura 19 – Criando uma VPC (3)
Fonte: Elaborado pelo autor (2024)

A console criará sua VPC. Confira as configurações e veja o status dela:

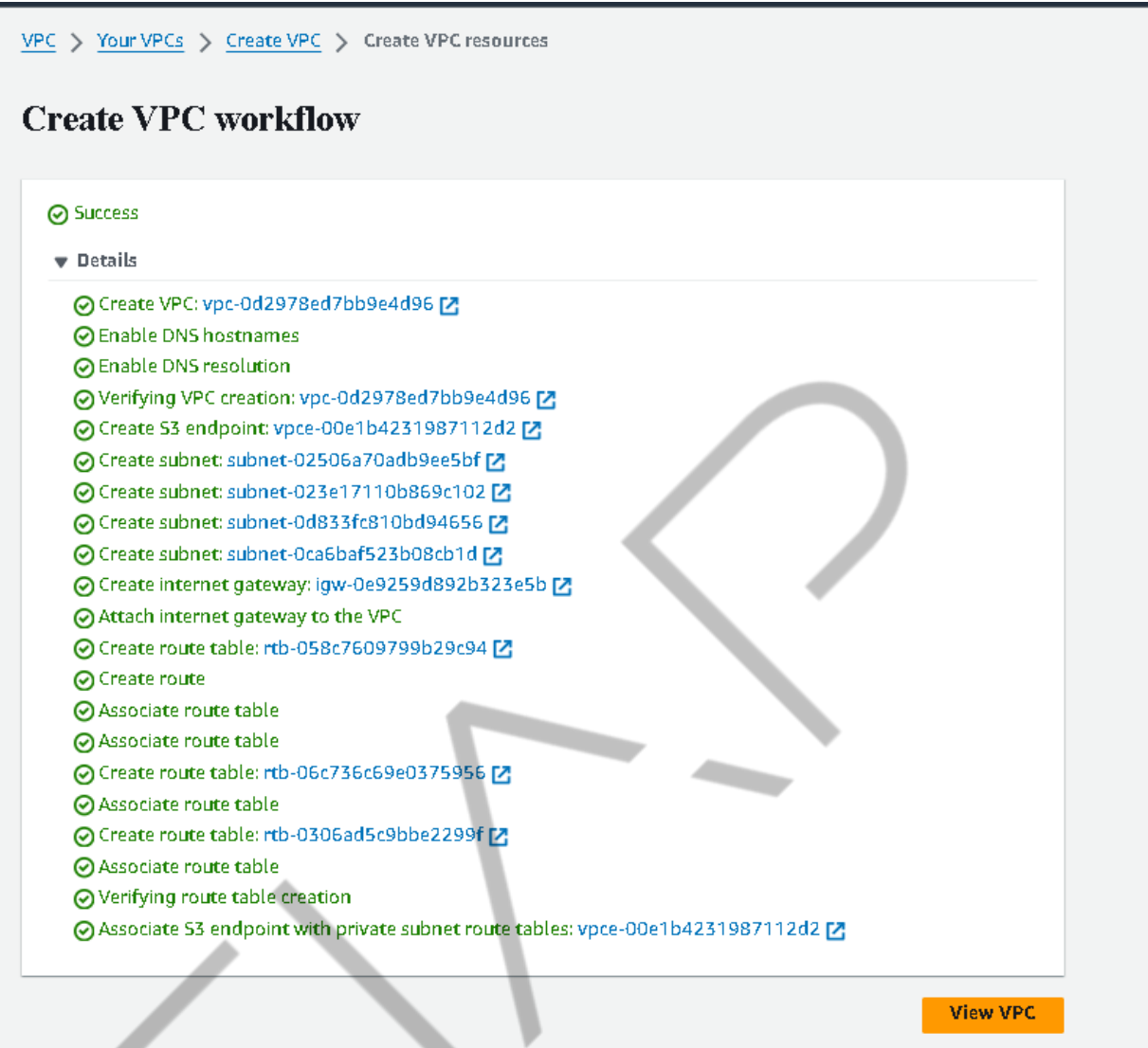


Figura 20 – Configurações da VPC
Fonte: Elaborado pelo autor (2024)

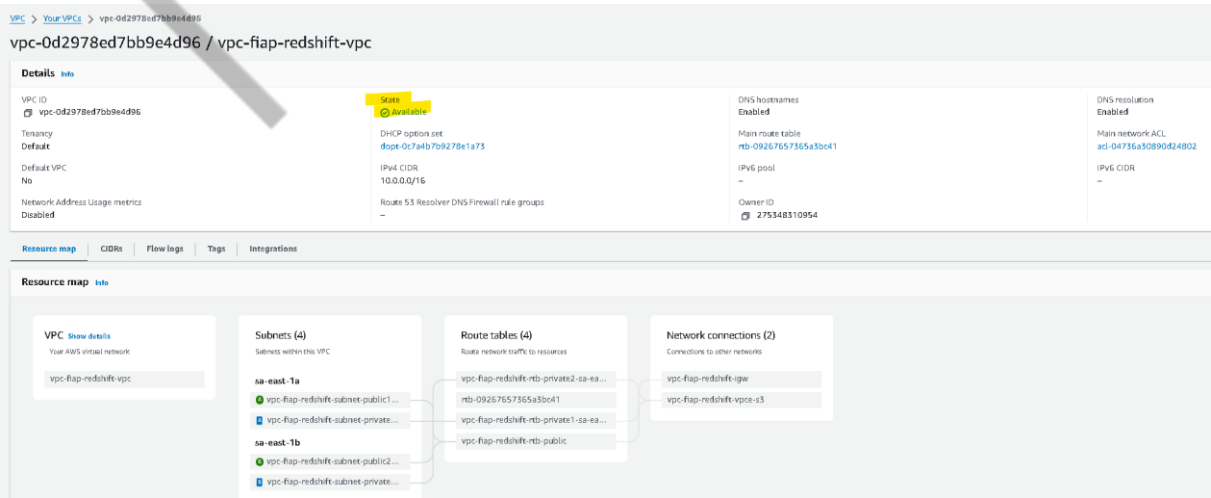


Figura 21 – Status da VPC
Fonte: Elaborado pelo autor (2024)

Subnet

Ainda no serviço de VPCs, encontre a opção de “subnets” no menu ao lado esquerdo:

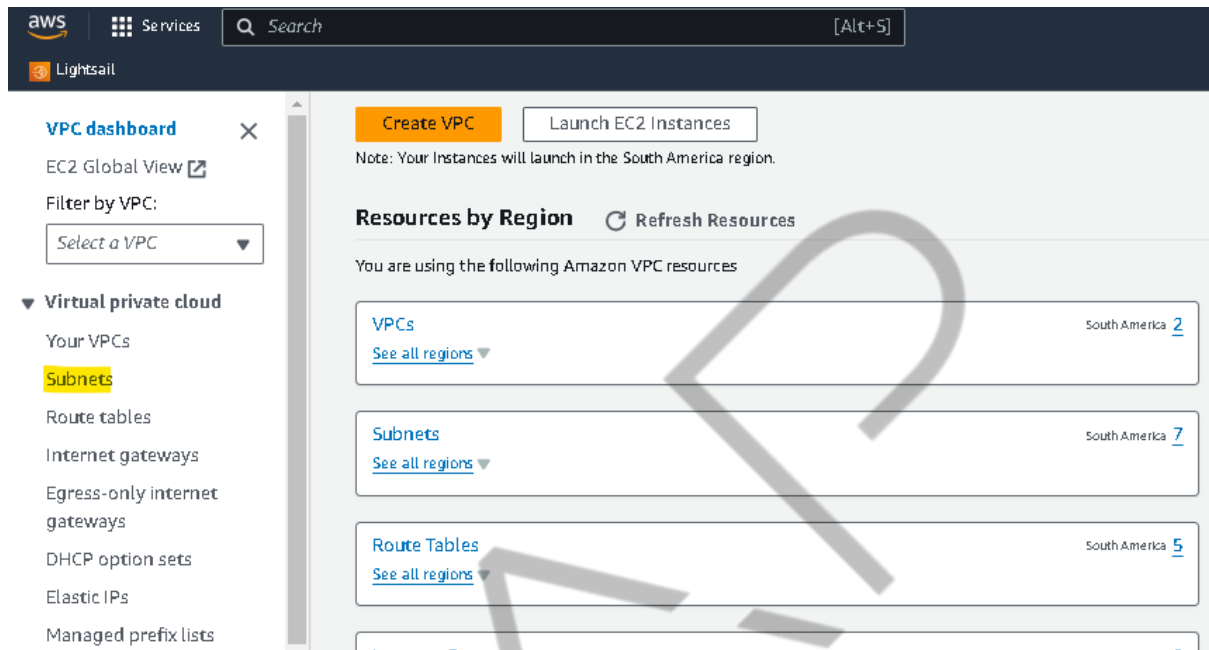


Figura 22 – Subnets
Fonte: Elaborado pelo autor (2024)

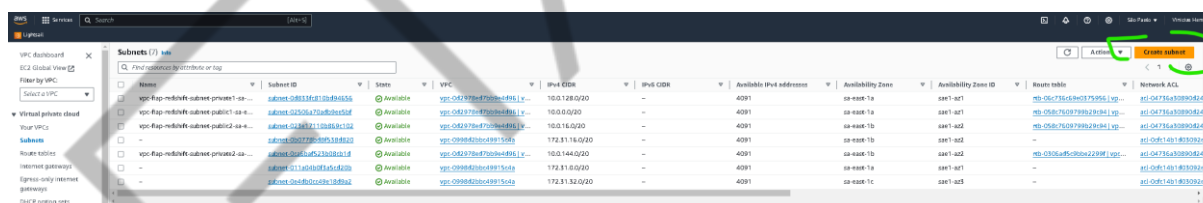
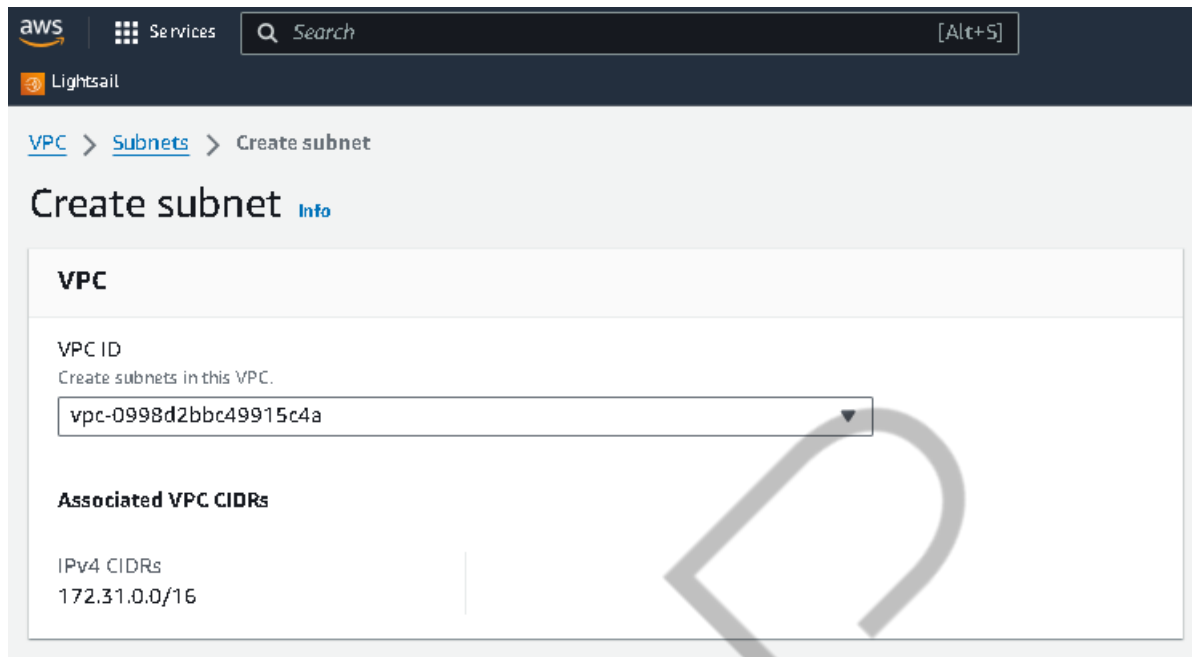


Figura 23 – Criando uma Subnet
Fonte: Elaborado pelo autor (2024)

Clique em “Create subnet”.

Selecione a VPC onde você deseja criar a Subnet. Caso você já tenha alguma VPC na região, pode seguir com ela; caso você tenha acabado de criar a VPC do passo anterior, selecione ela nessa configuração:



aws Services Search [Alt+S]

Lightsail

VPC > Subnets > Create subnet

Create subnet [Info](#)

VPC

VPC ID
Create subnets in this VPC.

vpc-0998d2bbc49915c4a

Associated VPC CIDRs

IPv4 CIDRs
172.31.0.0/16

Figura 24 – Configuração da Subnet
Fonte: Elaborado pelo autor (2024)

Security groups

Abra o serviço do EC2 e encontre a opções de “security groups”:

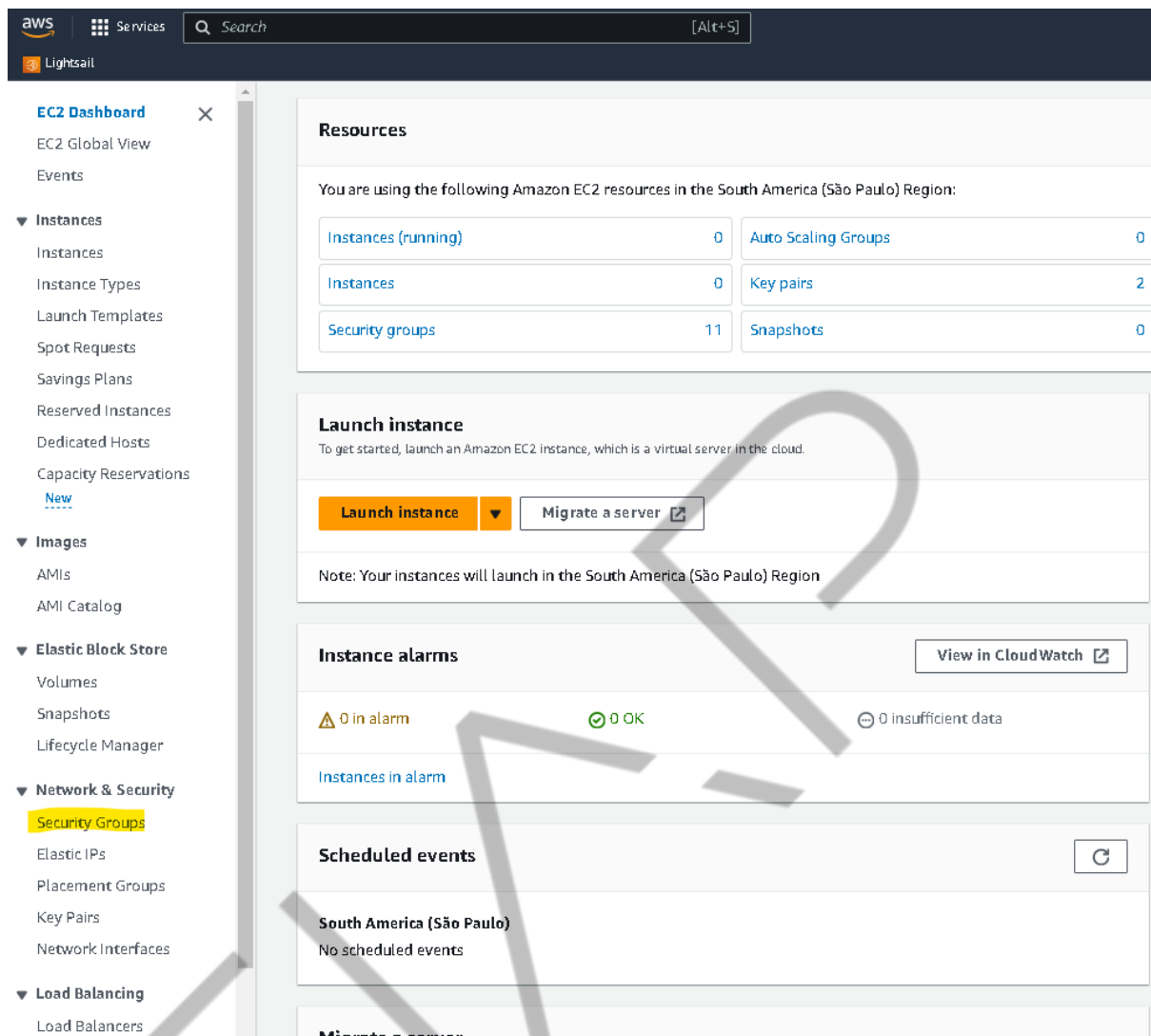


Figura 25 – Opção “Security Groups”
Fonte: Elaborado pelo autor (2024)

Clique em “Create security group”:

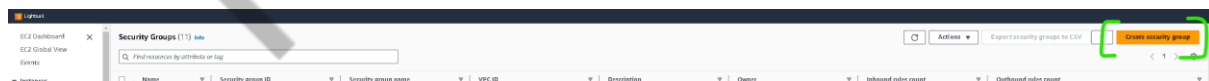


Figura 26 – Criando um Security Group (1)
Fonte: Elaborado pelo autor (2024)

EC2 > Security Groups > Create security group

Create security group [Info](#)

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic. To create a new security group, complete the fields below.

Basic details

Security group name [Info](#)

sgRedshift

Name cannot be edited after creation.

Description [Info](#)

sg para lab redshift

VPC [Info](#)

vpc-0d2978ed7bb9e4d96 (vpc-fiap-redshift-vpc)

Figura 27 – Criando um Security Group (2)
Fonte: Elaborado pelo autor (2024)

No restante das opções mantenha a configuração “padrão” e avance.

Security group sg-022b78bb85275f19a (sgRedshift) was created successfully

Details

sg-022b78bb85275f19a - sgRedshift

Details

| | | | |
|-----------------------------------|---|--|---------------------------------|
| Security group name sgRedshift | Security group ID sg-022b78bb85275f19a | Description sg para lab redshift | VPC ID vpc-0d2978ed7bb9e4d96 |
| Owner 27548310954 | Inbound rules count 0 Permission entries | Outbound rules count 1 Permission entry | |

Inbound rules

Search

| Name | Security group rule... | IP version | Type | Protocol | Port range | Source | Description |
|-------------------------------|------------------------|------------|------|----------|------------|--------|-------------|
| No security group rules found | | | | | | | |

Figura 28 – Configuração de Security Group
Fonte: Elaborado pelo autor (2024)

O QUE VOCÊ VIU NESTA AULA?

Nessa aula, você entendeu a diferença entre um sistema de dados voltado para operações transacionais e um sistema de dados voltado para trabalhos analíticos. Você também compreendeu os conceitos de “fatos” e “dimensões” por trás de um Data Warehouse e conseguiu aplicar isso tecnicamente.

Além disso, foi capaz de configurar e instanciar um cluster do Amazon Redshift para receber informações e ajudar nas tarefas de análises usando SQL, além de construir tabelas e consultas.

Por fim, também conseguiu trazer respostas sobre os dados usando funções de agregação de dados como somas ou contagens, além de trazer significado aos dados através dos joins entre fatos e dimensões.

REFERÊNCIAS

AWS. **Amazon Redshift best practices.** 2024. Disponível em: <<https://docs.aws.amazon.com/redshift/latest/dg/best-practices.html>>. Acesso em: 23 abr. 2024.

AWS. **Columnar storage.** 2024. Disponível em: <https://docs.aws.amazon.com/redshift/latest/dg/c_columnar_storage_disk_mem_mgmt.html>. Acesso em: 16 mai. 2024.

AWS. **Data warehouse system architecture.** 2024. Disponível em: <https://docs.aws.amazon.com/redshift/latest/dg/c_high_level_system_architecture.html>. Acesso em: 23 abr. 2024.

AWS. **SQL reference.** 2024. Disponível em: <https://docs.aws.amazon.com/redshift/latest/dg/cm_chap_SQLCommandRef.html>. Acesso em: 23 abr. 2024.

AWS. **Using a COPY command to load data.** 2024. Disponível em: <https://docs.aws.amazon.com/redshift/latest/dg/t>Loading_tables_with_the_COPY_command.html>. Acesso em: 23 abr. 2024.

PALAVRAS-CHAVE

Palavras-chave: DW. Data Warehouse. DM. Data Marts. Redshift. SQL. OLAP. OLTP. Fato. Dimensão.

EMENDAS



POSTECH