



BIG DATA

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA

INTEGRATION

Autores:

Vinícius Fernandes;

Rodrigo Martins;

Rodrigo Campos;

Michel Mansur;

Lucas Pereira;

Eric Tachdjian..

Data de criação: 24 de Outubro de 2023

SÃO PAULO – SP

2023

Controle de Documento

Histórico de Revisões

Table 1: Controle de documento

Data	Autor	Versão	Resumo da Atividade
24/10/23	Eric Tachdjian	1.0	Introdução (1.0/1.1)
24/10/23	Michel Mansur	1.1	Definição do problema (1.2)
25/10/23	Vinícius Fernandes	1.2	Proposta de Valor (3.1); Matriz de Risco (3.2); Análise de Tamanho de Mercado (3.3).
25/10/2023	Michel Mansur	1.3	Matriz de Risco (3.2) Persona (4.1)
25/10/23	Eric Tachdjian	1.4	User Story (4.3)

1. Introdução

Na dinâmica atual do mundo empresarial, a habilidade de gerenciar vastos volumes de dados e transformá-los em informações críticas é essencial para o êxito das organizações. Este projeto se propõe a criar um pipeline de Big Data, aproveitando os recursos poderosos disponibilizados pela Amazon Web Services (AWS). Seu propósito fundamental é capacitar as empresas a conduzirem análises estatísticas sobre dados armazenados em data lakes provisionado ou data warehouses (de acordo com a arquitetura proposta), permitindo a extração de insights fundamentais para embasar decisões estratégicas.

Para proporcionar isso, este projeto inclui a criação de um infográfico que proporcionará uma visualização mais intuitiva dos dados, simplificando a análise e potencializando a eficiência no processo de tomada de decisões.

1.1 Parceiro de Negócios

A Integration Consultoria é uma empresa especializada em consultoria estratégica e de gestão, com sua sede localizada no Brasil e presença em diversos países, incluindo Argentina, Chile, México, Estados Unidos, Reino Unido e Alemanha. Seu time de profissionais é capacitado para prestar suporte em diversas áreas, abrangendo Marketing e Vendas, Finanças e Administração, Tecnologia e Transformação Digital, Logística e Cadeia de Suprimentos, Sustentabilidade, bem como Implementação de estratégias.

1.2 Definição do Problema

1.2.1 Problema

O parceiro (Integration) manifestou a necessidade de uma ferramenta que lhe permitisse avaliar com precisão o potencial de consumo de cada categoria detalhadamente de seus clientes, nesse caso, levando em consideração variáveis como geografia (cidade) e canais de atendimento de cada categoria. Isso, por sua vez, permitirá o estabelecimento de metas estratégicas e implementação de táticas voltadas ao desenvolvimento de categorias específicas ou canais de distribuição específicos.

Para atender a esse requisito, o objetivo do cliente era ter um banco de dados robusto contendo as informações necessárias para uma análise precisa. Esta base deve ser acompanhada por uma representação visual que forneça os insights necessários para tomar decisões informadas durante as operações diárias.

2. Objetivos

2.1 Objetivos Gerais

Este projeto tem como objetivo o desenvolvimento de uma ferramenta com um alto ganho analítico e de inferência para outras áreas, baseados numa arquitetura robusta e de alta interoperabilidade. É esperada a criação de um pipeline de Big Data, utilizando a infraestrutura em cloud, para o gerenciamento de dados. O foco central reside na necessidade da criação de análises estatísticas detalhadas sobre áreas de foco do cliente, incluindo geografia e canais de atendimento.

Consequentemente, o resultado final será um sistema que permite uma avaliação precisa do potencial de consumo em categorias específicas, possibilitando que os vendedores estabeleçam metas estratégicas e implementem ações táticas direcionadas ao desenvolvimento de categorias ou canais de distribuição. Além disso, o projeto visa criar um infográfico adaptável que simplifique a análise, facilitando assim o processo de tomada de decisões informadas e estratégicas.

2.2 Objetivos Específicos

O projeto visa alcançar alguns objetivos específicos para atender às necessidades pontuais do cliente. Primeiramente, será desenvolvido um pipeline de Big Data, aproveitando as tecnologias avançadas da Amazon Web Services (AWS), para gerenciar grandes volumes de dados de forma eficiente e segura. Esse pipeline vai ser embarcado

num fluxo cíclico de dados, que virá de três fontes diferentes, sendo elas: as informações do cliente provisionadas pela API da consultoria, dados populacionais extraídos pelo órgão de pesquisa IBGE e outras fontes públicas.

A análise estatística detalhada será uma prioridade, incorporando variáveis como geografia (cidade) e canais de atendimento de cada categoria e tirando correlações e níveis de dependência das mesmas. Sendo assim, isso permitirá uma avaliação mais precisa do potencial de consumo, capacitando os vendedores com dados específicos para estabelecer metas estratégicas bem fundamentadas (com tomadas de decisão mais assertivas). Além disso, o projeto tem como objetivo viabilizar a implementação de ações táticas direcionadas ao desenvolvimento de categorias específicas ou canais de distribuição. Essas ações serão baseadas em insights derivados das análises feitas, permitindo uma abordagem “data-driven” para estratégias de vendas.

Por fim, a parte final e essencial deste projeto é a criação de um infográfico intuitivo. Esse infográfico proporcionará uma visualização cativante dos dados, simplificando a interpretação e, consequentemente, potencializando a eficiência no processo de tomada de decisões informadas e estratégicas durante as operações diárias da empresa.

2.3 Justificativa

A proposta de solução do nosso grupo é construir uma solução que entregue em um curto período uma análise confiável a partir dos dados inseridos pelo usuário. Esta é a ideia primária na solução, uma vez que é essencial para o profissional de marketing e para o analista de dados que as informações geradas a partir dos dados possibilitem uma formulação prática de insights e de novas estratégias.

Além disso, pretendemos entregar um sistema que permita flexibilidade dos inputs, adaptando-se a diferentes necessidades do cliente. Desta forma, os resultados obtidos para cenários diversos serão úteis, independente da variação da situação, trazendo um maior potencial à solução final.

É importante destacar que o sistema entregue deve permitir um tráfego seguro dos dados inseridos. Por fim, com o infográfico, a entrega final possibilitará a geração de insights a partir da integração efetiva entre as tecnologias utilizadas na arquitetura planejada, de forma segura e eficiente.

3. Compreensão do Problema

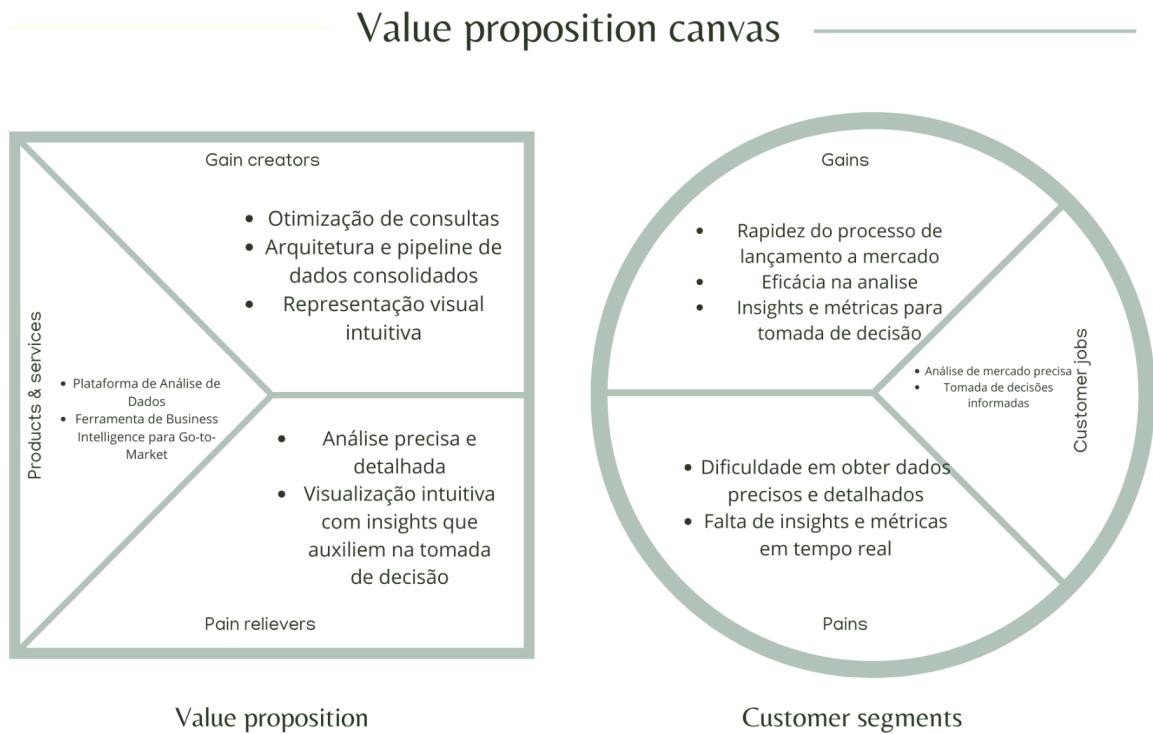
O problema em questão envolve um cliente distribuidor, da Integration, que opera em diversas categorias nos setores alimentar e de food service. Esse cliente busca uma ferramenta que o auxilie a entender o potencial de consumo de cada categoria em níveis detalhados, como cidade e canal de atendimento para cada categoria específica. O objetivo é possibilitar o cliente a tomar decisões estratégicas, como direcionar sua equipe de vendas, implementar ações táticas para promover categorias ou canais específicos e, para isso, ele precisa de informações detalhadas disponíveis em uma base de dados que possibilite análises, idealmente com visualizações que facilitem a tomada de decisões no dia a dia de suas operações.

3.1 Proposta de Valor

A proposta de valor no marketing destaca um negócio, posicionando-o com uma visão estratégica, e mais direcionada para seu público. Esse tipo de análise reforça a habilidade da empresa em entender as necessidades dos clientes, com uma visão mais focada nos mesmos. Sendo assim, para sua melhor visualização, podemos utilizar algumas ferramentas.

O Canvas Proposta de Valor é uma dessas ferramentas que é muito utilizada no mercado e alinha os produtos ou serviços de uma empresa com as necessidades dos clientes. Ele destaca o segmento de clientes, a proposta de valor exclusiva, os produtos oferecidos, as preocupações e benefícios dos clientes, e como a solução resolve esses problemas e proporciona ganhos. Com esta visão, e pensando no objetivo, nas

necessidade do cliente e possíveis ganhos com o projeto, determinamos o Canvas para o projeto:



Informações do Canvas Value Proposition:

- **Value Proposition**
 - *Products & Services:*
 - Plataforma de análise de dados
 - Ferramenta de Business Intelligence para Go-to-Market
 - *Gain creators:*
 - Otimização de consultas
 - Arquitetura e pipeline de dados consolidados
 - Representação visual intuitiva
 - *Pain relievers:*
 - Análise precisa de detalhada
 - Visualização intuitiva com insights que auxiliem na tomada de decisão

- **Customer Segments**

- *Customer jobs:*

- Análise de mercado precisa
 - Tomada de decisões informadas

- *Gains:*

- Rapidez do processo de lançamento a mercado
 - Eficácia na análise
 - Insights e métricas para tomada de decisão

- *Pains:*

- Dificuldade em obter dados precisos e detalhados
 - Falta de insights e métricas em tempo real

3.2 Matriz de Risco

A matriz de risco pode ser definida como uma ferramenta essencial utilizada para identificar e avaliar os possíveis riscos que podem impactar um projeto. Ela classifica os riscos em categorias com base na probabilidade de ocorrência e no impacto potencial, permitindo às equipes priorizá-los e criar estratégias de mitigação eficazes. Ao representar graficamente os riscos em uma matriz, a equipe pode tomar decisões mais bem fundamentadas sobre onde direcionar recursos e esforços, encontrando um equilíbrio entre os riscos e as oportunidades. Isso promove uma abordagem proativa para gerenciar incertezas e superar possíveis obstáculos. Segue abaixo matriz de risco desenvolvida pela equipe em visão dos possíveis riscos (juntamente com plano de ação)

e oportunidades em relação ao projeto:

Ameaças							Oportunidades			
	Ameaça 8: Dependência de terceiros, como fornecedores da nuvem	-	-	-	-	-	Oportunidade 3: Desenvolvimento de infográficos, com dados em tempo real, e previsão analítica do consumo nas categorias pré-definidas	-	-	-
90%	-	-	-	Ameaça 7: Problemas relacionados às contas aws	Ameaça 1: Ausência de um integrante do time	Oportunidade 1: Criação de ferramenta que auxilie na tomada de decisão dos cliente	Oportunidade 2: Realização de testes que comprove uma alta precisão da ferramenta	Oportunidade 6: Fácil portabilidade (em termos de funcionalidade e serviços) de serviços Azzure para arquitetura em AWS	-	-
70%	-	-	-	Ameaça 6: O cliente pode não adotar ou usar efetivamente o infográfico gerado.	Ameaça 2: Alta expectativa do cliente e abstração do escopo do projeto	Ameaça 3: Criação de infográficos que não auxiliam na tomada de decisão ao cliente	-	Oportunidade 4: Criação de um pipeline de dados bem estruturado	-	-
50%	Ameaça 5: Dados inconclusivos, massivos ou de baixa qualidade podem prejudicar a análise estatística.	Ameaça 6: O cliente pode não adotar ou usar efetivamente o infográfico gerado.	Ameaça 4: Baixa granularidade e insights enviesados (sem embasamento do cliente) na análise estatística	-	-	-	Oportunidade 5: Redução de custo no uso de uma infraestrutura cloud linear, separando serviços em diferentes clouds	-	-	-
30%	-	-	-	-	-	-	-	-	-	-
10%	-	-	-	-	-	-	-	-	-	-
	Muito baixo	Baixo	Moderado	Alto	Muito alto	Muito alto	Alto	Moderado	Baixo	Muito Baixo
	Impacto									

3.2.1 Ameaças e Plano de Ação

Segue descrição de todas as ameaças denotadas na Matriz de Risco e seus respectivos planos de ação:

Ameaça 01: Ausência de um integrante do time

- **Plano de Ação:** Redistribuir as tarefas atribuídas ao integrante, e garantir que a qualidade do projeto não diminua.
- **Responsável:** Grupo

- **Justificativa:** Devido a problemas familiares, um integrante do grupo não conseguiu participar da sprint, fazendo com que houvesse a necessidade de redistribuição de suas tarefas com o restante do grupo.
- **Risco:** 70%

Ameaça 02: Alta expectativa do cliente e abstração do escopo do projeto.

- **Plano de Ação:** Assegurar a presença do cliente desde o início do projeto para alinhar as expectativas com a realidade
- **Responsável:** Michel
- **Justificativa:** À medida que o escopo do projeto se expande devido a expectativas não gerenciadas, os prazos estabelecidos inicialmente podem não ser cumpridos.
- **Risco:** 30%

Ameaça 03: Criação de infográficos que não auxiliam na tomada de decisão ao cliente

- **Plano de Ação:** Validação recorrente com o cliente, para o desenvolvimento de visualizações com métricas que realmente auxiliam na tomada de decisão.
- **Responsável:** Vinícius
- **Justificativa:** A criação de infográficos ineficazes não contribuem efetivamente para a tomada de decisões dos clientes, trazendo várias consequências negativas,

como decisões tomadas de forma errada, e perda de confiança na utilização da ferramenta.

- **Risco:** 30%

Ameaça 04: Baixa granularidade e insights enviesados (sem embasamento do cliente) na análise estatística

- **Plano de Ação:** Pesquisa assertiva e testes constante para que não haja erro nos resultados da análise estatística
- **Responsável:** Michel
- **Justificativa:** Desacredita e invalida qualquer inferência feita a partir da análise, assim como a perda de personalização dos dados, e visualizações estatísticas
- **Risco:** 30%

Ameaça 05: Dados inconclusivos, massivos ou de baixa qualidade podem prejudicar a análise estatística.

- **Plano de Ação:** Realizar uma avaliação da infraestrutura atual identificando pontos fracos e gargalos, definindo o que pode ser feito.
- **Responsável:** Rodrigo Martins
- **Justificativa:** Dados de baixa qualidade podem levar a resultados estatísticos incorretos, prejudicando a tomada de decisões com base nas análises.
- **Risco:** 50%

Ameaça 06: O cliente pode não adotar ou usar efetivamente o infográfico gerado.

- **Plano de Ação:** Colaborar estreitamente com o cliente ao longo do projeto para entender suas necessidades e preferências.
- **Responsável:** Rodrigo Campos
- **Justificativa:** Os recursos investidos na criação do infográfico podem ser desperdiçados se ele não
- **Risco:** 50%

Ameaça 07: Problemas relacionados às contas da AWS.

- **Plano de Ação:** Confirmar a melhor maneira de “atacar” esse ponto com o professor responsável e garantir a existências de backups fora da AWS.
- **Responsável:** Vinícius Fernandes
- **Justificativa:** Após um determinado tempo de utilização, as contas utilizadas na AWS são automaticamente excluídas e todo progresso realizado dentro delas são perdidos.
- **Risco:** 70%

Ameaça 08: Dependência de serviços terceiros, como fornecedores da nuvem

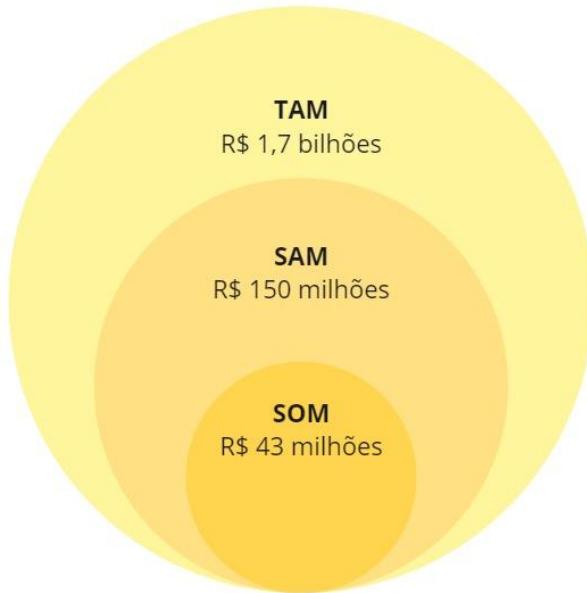
- **Plano de Ação:** Validar os "commits" (principalmente na main) e validar com o cliente se há necessidade de mascarar dados ou realizar anonimização.

- **Responsável:** Lucas
- **Justificativa:** A dependência de terceiros faz com que o sistema seja vulnerável, tenha falta de controle e inclua riscos financeiros.
- **Risco:** 70%

3.3 Análise de Tamanho de Mercado

Com a intenção de entendimento do público alvo e impacto da solução aos clientes no longo prazo, podemos usar algumas estratégias para mensurar o tamanho do seu mercado, entre elas o TAM, SAM, SOM. Esse tipo de análise de mercado, representa diferentes subconjuntos de um mercado, na qual há uma previsão em relação a demanda sobre os produtos ou serviços, projetando sua performance, baseando-se em premissas de pesquisas segmentadas e com recortes pré-estabelecidos de forma estratégica.

Direcionando essa análise para a realidade do cliente, partimos do entendimento de que a solução proposta atende a um mercado específico (no caso, o de Business Intelligence), com projeção de curto, médio e longo prazo. Segue análise e tamanho de mercado em conforme pesquisa aprofundada:



TAM: Total Addressable Market

Trata-se da soma da receita de todas as empresas de um segmento e pode incluir também organizações que comercializam soluções alternativas, mas que também são tidas como concorrentes dentro de um determinado mercado.

SAM: Serviceable Available Market

Parcela do mercado que uma organização pode alcançar em um futuro próximo (em média, 5 anos), de acordo com seus recursos, sendo muito útil para projetar o crescimento da empresa.

SOM: Serviceable Obtainable Market

Traz uma perspectiva mais realista sobre qual parcela do mercado uma organização pode conquistar, com base no momento atual do negócio. Esse indicador deve ser utilizado para **perspectivas de curto prazo** (de 1 a 2 anos).

TAM:
Mercado de
Business
Intelligence

SAM:
Segmento de
varejo alimentar e
de serviço
alimentar no
Brasil

SOM:
% do varejo
alimentar
em São Paulo

3.3.1 Definição TAM, SAM, SOM:

3.3.1.1 TAM - Total Available Market ou Mercado Total

Refere-se à procura integral do mercado por um produto ou serviço, sendo feito um recorte por segmento mais amplo, com uma visão estratégica ao cliente. Normalmente, calcula-se somando as receitas do mercado em análise. Com a visão analítica de com base na pesquisa de mercado direcionada ao segmento de Business Intelligence, obtivemos a receita aproximada de R\$1,7 bilhão de reais.

3.3.1.2 SAM - Serviceable Available Market ou “Mercado Endereçável”

Determina a projeção na qual uma organização pode alcançar algum segmento específico em um futuro próximo (em média, 5 anos), de acordo com seus recursos, com base em pesquisas de mercado no segmento de varejo alimentar e de serviço alimentar no Brasil (recorte geográfico). A receita aproximada neste quesito é de R\$150 milhões de reais.

3.3.1.3 SOM - Serviceable Obtainable Market ou “Mercado Acessível”

Pode ser definida como a parte mais “nichada” da análise. Essa análise traz uma perspectiva mais realista sobre qual parcela do mercado na qual há possibilidade de conquistar, com base no momento atual do negócio. Esse indicador é utilizado para perspectivas de curto prazo (de 1 a 2 anos), por isso possui um recorte mais segmentado dentro da proposta da empresa. De acordo com o projeto, e a demanda apresentada, realizamos o recorte no segmento de distribuição alimentar, com base no estado de São Paulo, retirando a porcentagem de empresas que já utilizam Business Intelligence.

4. Análise de Experiência do Usuário

4.1 Personas

A Persona é a representação fictícia de um cliente alvo (do time de desenvolvimento), baseada em dados demográficos, comportamentais, necessidades e desejos. Ela ajuda a compreender e focar nas características-chave do público-alvo, permitindo a criação de estratégias e desenvolvimento de produtos/serviços mais direcionados e eficazes. A persona é uma ferramenta valiosa para entender quem são os clientes e como atendê-los de maneira personalizada, proporcionando uma experiência, com o sistema, mais relevante e satisfatória.

No projeto em questão, as personas são representadas por dois profissionais de mercado: Ricardo Silva, cientista de dados, e Lucas Macedo, consultor de marketing. Ricardo e Lucas personificam os clientes alvos.. Com base em suas características e necessidades, as personas de Ricardo e Lucas serão fundamentais para nos orientar durante o processo de desenvolvimento do projeto.

4.1.1 Ricardo Silva - Cientista de Dados

Ricardo Silva é um Cientista de Dados apaixonado por transformar informações em conhecimento. Com formação em Estatística e experiência em diversos setores, ele se encontrou na área de Dados, e atua majoritariamente na manutenção de algoritmos. Sua expertise técnica e habilidade em identificar padrões e tendências relevantes permitem que tome decisões estratégicas de acordo com o processo anterior de engenharia de dados. Além disso, sua comunicação eficaz e foco em resultados o tornam capaz de traduzir análises complexas em informações comprehensíveis para que haja conhecimento nos dados inicialmente coletados. Ricardo está comprometido em fornecer insights açãoáveis e contribuir para o sucesso do projeto em questão.

RICARDO SILVA

CIENTISTA DE DADOS



BIOGRAFIA

NOME: Ricardo Silva

IDADE: 32 anos

GÊNERO: Masculino

LOCALIZAÇÃO: São Paulo, Brasil

RENDA: Por volta de R\$ 9 mil

OCUPAÇÃO: Cientista de dados

CARACTERÍSTICAS

- Analítico;
- Orientado a resultados;
- Ambicioso;
- Atento.

DORES

Ricardo atualmente sofre com alguns projetos e sistemas que caem em sua mão, normalmente vindos do time de engenharia, com muitos gargalos e falhas durante o processo. Isso vem antes da parte de desenvolvimento, mas sim na parte de pré-processamento e tratamento dos dados. Segue descrição das dores principais de Ricardo:

- Dificuldade de obter dados confiáveis;
- Complexidade na limpeza e preparação de dados
- Identificação de padrões e tendências relevantes
- Comunicação eficaz dos insights

FORMAÇÃO

Possui bacharelado em Ciência da Computação pela USP e mestrado em Estatística na PUC-SP

HÁBITOS RELACIONADOS AO SISTEMA:

Será responsável por agregar novas funcionalidades no sistema, especificamente para cada uma das áreas e stakeholders envolvidos

NECESSIDADES E DESEJOS

Ricardo Silva, Cientista de dados, busca atender suas necessidades e realizar seus desejos no campo da análise de dados. Ele busca acesso a dados confiáveis e bem estruturados, além de ferramentas eficientes para limpeza e preparação desses dados. Ricardo almeja aprimorar suas habilidades analíticas, identificando padrões e tendências relevantes, e deseja também a capacidade de comunicar insights de maneira clara e impactante. Além disso, ele busca estar sempre atualizado em tecnologias e técnicas de análise de dados, desejando recursos e treinamentos para alcançar esse objetivo. Segue descrição dos desejos e necessidades principais:

- Pipeline de dados estruturado para continuação do projeto
- Arquitetura consolidada, com recursos atentos aos 5V's de Big Data
- Olhar atento do time de engenharia no processamento dos dados

4.1.1.2 Cenários de interação no sistema e nível de letramento digital

Conforme a solução apresentada, entendemos que a partir da solução, Ricardo guiará os próximos passos a partir do pipeline estruturado. Sendo assim, ele terá interação direta com os atributos técnicos da solução. Segue abaixo mapa do cenário de atuação da persona, Ricardo Silva:

RICARDO SILVA

Cientista de Dados



↑ Desenvolvimento de melhorias e próximos passos



↓ Sabe o conceito, mas não atua na inferência de métricas



↑ Manutenção ativa da ferramenta



Possui entendimento completo da arquitetura do sistema e seus gargálos



“Precisamos fazer uma automatização dos logs de venda para o time de marketing ter este controle específico.”

4.1.2 Lucas Macedo - Consultor de Marketing

Já o Lucas Macedo, é consultor de marketing apaixonado por estratégias criativas e inovação, traz consigo uma sólida carreira dedicada a impulsionar marcas e maximizar a visibilidade online. Reconhecido por seu talento em criar campanhas eficazes, Lucas busca constantemente novas maneiras de conectar marcas ao seu público, combinando sua criatividade inigualável com análises detalhadas de dados para atingir resultados excepcionais.

LUCAS MACEDO

CONSULTOR DE MARKETING



NOME: Lucas Macedo

IDADE: 46 anos

GÊNERO: Masculino

LOCALIZAÇÃO: São Paulo, Brasil

RENDA: Acima de R\$ 10 mil

OCUPAÇÃO: Consultor de Marketing e Vendas

BIOGRAFIA

Lucas Macedo, consultor de marketing apaixonado por estratégias criativas e inovação, traz consigo uma sólida carreira dedicada a impulsionar marcas e maximizar a visibilidade online. Reconhecido por seu talento em criar campanhas eficazes, Lucas busca constantemente novas maneiras de conectar marcas ao seu público, combinando sua criatividade inigualável com análises detalhadas de dados para atingir resultados excepcionais.

CARACTERÍSTICAS

- Inovador;
- Resiliente;
- Cauteloso;
- Atento.

DORES

Lucas Macedo enfrenta dores relacionadas à eficiência de processos em sua rotina. Ele acredita que muitos procedimentos poderiam ser mais eficazes e que a automação é fundamental para otimizar suas tarefas. Além disso, a organização pessoal também se apresenta como um desafio constante para ele. Segue descrição das dores principais de Lucas:

- Processos de lançamento poderiam ser mais eficientes;
- Setup inicial para lançamento deve ser automatizado
- Organização e processo lento de levantamento dos dados para lançamentos a mercado
- Dificuldade na Tomada de Decisão Informada

FORMAÇÃO

Ensino Superior em Publicidade e Propaganda pela ESPM. E pós-graduação na universidade de Westminster

HÁBITOS RELACIONADOS AO SISTEMA:

Utilizará a ferramenta em seu estado final, no qual retirará insights e métricas relavantes para tomada de decisão

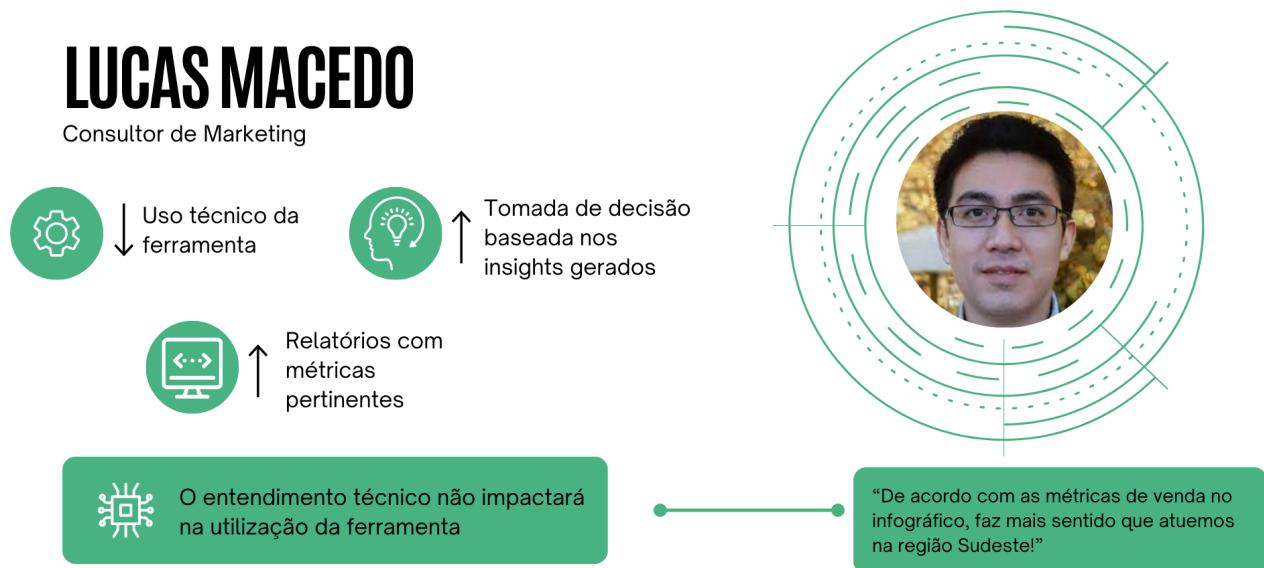
NECESSIDADES E DESEJOS

Para atender às necessidades de Lucas, é fundamental implementar um sistema mais robusto que simplifique e agilize seus processos. Isso envolve a automação de tarefas e a otimização de fluxos de trabalho. Além disso, a capacidade de analisar e consultar dados com maior facilidade é crucial para tomar decisões informadas e melhorar a eficiência de suas atividades diárias. Lucas busca uma solução que atenda a esses requisitos, permitindo-lhe alcançar seus objetivos com mais eficácia. Segue descrição dos desejos e necessidades principais:

- Sistema mais robusto, que facilita os processos de lançamento desde o início
- Analisar e consultar os dados com maior facilidade
- Metas Estratégicas e Ações Táticas de lançamento
- Representação Visual de Dados

4.1.2.2 Cenários de interação no sistema e nível de letramento digital

No caso do Lucas, é de se esperar que sua interação diante da solução seja muito mais baseada em tomadas de decisão assertivas de acordo com os insights gerados pela ferramenta. Deste modo, a interação será diretamente com a solução final, na representação visual e inferências de negócios, trazidas pelo infográfico disponibilizado e alimentadas pelo pipeline de dados criado. Segue ilustração do cenário de interação da persona, Lucas Macedo:



4.2 Jornada do Usuário

A jornada do usuário consiste no caminho que uma pessoa percorre ao interagir com um produto, serviço ou sistema. É uma história que descreve as experiências, emoções e ações de um usuário ao longo do tempo, o mapa que mostra todas as etapas que alguém percorre ao usar algo. Compreender a jornada do usuário ajuda a melhorar a experiência geral, identificando pontos de dor e oportunidades de aprimoramento.

4.2.1 Ricardo Silva - Analista de Dados

Ricardo Silva, o analista, está focado em utilizar ferramentas para atender às demandas de coleta, limpeza e análise de dados complexos. Ele busca otimizar a eficácia de suas análises e transformar insights em decisões estratégicas bem embasadas, com o objetivo de impulsionar o sucesso dos projetos nos quais está envolvido.

4.2.2 Expectativas

Ricardo tem expectativas claras quanto à solução de análise de dados que utiliza. Ele anseia por dados precisos e confiáveis sobre o mercado e o público-alvo a fim de embasar suas decisões estratégicas. Além disso, espera que a ferramenta otimize o processo de análise, contribua para a eficiência de suas tarefas e traduza insights complexos em ações práticas, impulsionando resultados mensuráveis em seus projetos.

4.2.3 Oportunidades

- Há oportunidades para melhorar a apresentação dos dados, garantindo que Ricardo tenha acesso claro às funcionalidades e benefícios da solução.
- Aprimoramentos na interface da solução podem permitir análises mais detalhadas e resultar em um investimento mais preciso de recursos.

4.2.4 Responsabilidades

- Utilizar a solução para analisar os dados de mercado.
- Assegurar a segurança das informações obtidas ao utilizar os dados do cliente.
- Interpretar os resultados e aplicar os insights na estratégia de marketing.

- Realizar simulações da estratégia planejada, utilizando a solução de análise de dados como uma ferramenta contínua de otimização

4.2.5 Pensamentos

- Fase 1: "*Preciso acessar os dados para diagnosticar o problema do cliente da melhor maneira possível.*"
- Fase 2: "*Esta solução parece a mais adequada para atender às nossas necessidades, mas não tenho certeza se os resultados serão satisfatórios*"
- Fase 3: "*Ao utilizar a solução, sinto que estou investindo recursos, mas ao mesmo tempo posso simular cenários diversos com a ferramenta, extraindo insights valiosos para minhas análises*"
- Fase 4: "*Enfrentei as questões que impediam a praticidade das minhas análises. Com perseverança, consegui obter resultados significativos, desenvolvendo uma estratégia comercial eficaz com base nos insights extraídos da análise.*"

4.2.6 Sentimentos

- Fase 1: Inovador e Cauteloso
- Fase 2: Desconfiado
- Fase 3: Curioso e inseguro

- Fase 4: Satisfeito

- Fase 5: Satisfeito

4.2.7 Picos e Vales

- Fase 1: Pico → Nessa fase o Ricardo está esperançoso quanto a nova tecnologia que deseja adotar e vê grande potencial com seu uso.
- Fase 2: Vale → Nessa fase, Ricardo começou a utilizar a ferramenta, se complicou com algumas funcionalidades, mas começou a se acostumar com o sistema.
- Fase 3: Pico → Ricardo começou o uso efetivo e começou a sentir o real potencial da ferramenta, possibilitando uma análise muito mais completa.
- Fase 4: Pico → Ricardo começou a superar os primeiros obstáculos que tinha antes de adotar a ferramenta e se animou com a eficácia dos insights
- Fase 5: Pico → Os resultados começaram a aparecer e Ricardo sentiu o impacto que a nova ferramenta causou, impulsionando seu sucesso com as estratégias de mercado e diminuindo o tempo utilizado para as análises.

4.2.8 Touchpoints

- Ricardo Silva: acesso técnico à ferramenta, primeiramente com acesso e autorização ao ambiente da AWS, e provavelmente em uma IDE de desenvolvimento.

4.2.9 Jornada

FASE 1: Identificação das Necessidades	FASE 2: Adoção da Solução	FASE 3: Uso da Solução	FASE 4: Desafios Superados	FASE 5 Alcançando Resultados
<ul style="list-style-type: none"> Ricardo é um analista de dados inovador, sempre em busca de soluções eficazes para suas análises. Ele busca uma solução que forneça dados precisos e confiáveis sobre o mercado e o público-alvo do cliente. Ricardo busca praticidade e rapidez para suas análises. <p><i>"Preciso acessar os dados para diagnosticar o problema do cliente da melhor maneira possível."</i></p> <p>Sentimento: Inovador e Cauteloso</p>	<ul style="list-style-type: none"> Ricardo escolhe adotar uma solução de análise de dados hospedada na AWS, que reúne informações do cliente e dados públicos em um ambiente seguro. Essa escolha permite a análise do setor do cliente, informações demográficas e tendências de mercado, fornecendo insights valiosos. <p><i>"Esta solução parece a mais adequada para atender às nossas necessidades, mas não tenho certeza se os resultados serão satisfatórios"</i></p> <p>Sentimento: Desconfiado</p>	<ul style="list-style-type: none"> Apesar de gastar recursos ao utilizar a plataforma para gerar infográficos, Ricardo considera que a solução permite simular diferentes cenários e é eficaz em fornecer insights. A rapidez em que o sistema funciona e a praticidade em utilizá-lo auxiliam Ricardo em seu processo de análise. <p><i>"Ao utilizar a solução, sinto que estou investindo recursos, mas ao mesmo tempo posso simular cenários diversos com a ferramenta, extrair insights valiosos para minhas análises"</i></p> <p>Sentimento: Curioso e Inseguro</p>	<ul style="list-style-type: none"> Ricardo supera os desafios associados ao gasto de recursos e de tempo para gerar infográficos. Ele obtém resultados significativos, desenvolvendo uma estratégia comercial eficaz com base nos insights da análise. <p><i>"Enfrentei as questões que impediam a praticidade das minhas análises. Com perseverança, consegui obter resultados significativos, desenvolvendo uma estratégia comercial eficaz com base nos insights extraídos da análise."</i></p> <p>Sentimento: Satisfeito</p>	<ul style="list-style-type: none"> A análise é finalizada em um período de tempo reduzido em relação ao que seria gasto sem o uso do sistema; Ricardo consegue gerar um infográfico útil para a visualização das informações obtidas; O analista alcança sucesso na estratégia comercial, adaptando-a conforme necessário e finalizando o projeto com êxito. <p>Sentimento: Satisfeito</p>

4.2.10 Desfecho

- Ricardo possui entendimento e alto controle da ferramenta. É de se esperar que haja uma arquitetura consolidada, e que o fluxo de dados esteja funcionando, com o número de requisições que normalmente sejam recebidas

4.2.11 Insights

- Há possibilidade da criação de um sistema maior com mais fontes de entrada dos dados, e possibilidade de expansão do número de requisições abarcadas pelo sistema. Neste caso, há possibilidade de integração de mais serviços da AWS, e redução de custos para outros serviços “open source”.

4.2.2 Lucas Macedo - Consultor de Marketing e Vendas

4.2.2.1 Expectativas

Lucas reconhece a necessidade de otimizar seus processos de coleta de dados, assegurando que os dados coletados sejam precisos e relevantes para suas estratégias de Go-To-Market. Ele espera que o sistema ofereça visualizações de dados interativas e painéis personalizados que sejam intuitivos e fáceis de usar. Isso ajudará sua equipe a compreender os dados de forma clara e a colaborar na geração de ideias e decisões estratégicas.

4.2.2.2 Oportunidades

- Encontrar uma solução que integre efetivamente todas as etapas de lançamento, automação e análise.
- A automação reduzirá a carga de trabalho de Lucas, permitindo-lhe focar em estratégias criativas.

4.2.2.3 Responsabilidades

- Lucas é responsável por avaliar as opções disponíveis no mercado e tomar uma decisão informada sobre a solução a ser adotada.
- Desenvolver e implementar estratégias de Go-To-Market para novos produtos ou serviços, identificando público-alvo, canais de distribuição, mensagens chave e preços.
- Aconselhar sobre o desenvolvimento ou adaptação de produtos e serviços para atender às demandas do mercado.

4.2.2.4 Pensamentos

- Fase 1: "Preciso aprimorar minhas estratégias de Go-To-Market. As atuais não estão fornecendo os resultados desejados."
- Fase 2: ""Vamos fazer um investimento na implementação e treinamento para que todos possam aproveitar ao máximo a plataforma."
- Fase 3: "Esta solução facilita a análise avançada e torna nossas decisões mais informadas."
- Fase 4: "Estou me tornando mais proficiente na análise de dados à medida que superamos esses obstáculos."

4.2.2.5 Sentimentos

- Fase 1: Determinado
- Fase 2: Empolgado
- Fase 3: Satisfeito
- Fase 4: Confiante
- Fase 5: Muito Satisfeito e otimista

4.2.2.6 Picos e Vales

- Fase 1: Pico → Lucas percebeu uma oportunidade de melhora nas parte de decisões estratégicas e decide procurar uma solução para otimizar seus processos.

- Fase 2: Pico → Nessa fase, Lucas começou a pesquisar as diferentes ferramentas de para melhorar suas estratégias de go-to-market. Encontrou a da M&C Solutions e investiu.
- Fase 3: Vale → Nessa fase, Lucas começou a utilizar a ferramenta, enfrentou algumas dificuldades com algumas funcionalidades, se confundiu um pouco, mas se manteve aberto à ferramenta.
- Fase 4: Pico → Lucas começou a superar os obstáculos iniciais que enfrentou antes de adotar a ferramenta e ficou empolgado com a eficácia dos insights.
- Fase 5: Pico → Os resultados começaram a se manifestar, e Lucas experimentou o impacto positivo que a nova ferramenta trouxe, impulsionando seu sucesso nas estratégias de mercado e reduzindo o tempo necessário para as análises.

4.2.2.7 Touchpoints

- Acesso direto aos infográficos, por meio do Amazon Quicksight (ou outra ferramenta de visualização) e/ou aplicação web ("nice to have").

4.2.2.8 Jornada de Usuário

FASE 1: Identificação das Necessidades	FASE 2: Adoção da Solução	FASE 3: Uso da Solução	FASE 4: Desafios Superados	FASE 5 Alcançando Resultados
<ul style="list-style-type: none"> Lucas Macedo, consultor de marketing, reconhece a necessidade de aprimorar suas estratégias de go-to-market. Ele busca uma solução que otimize seus processos de lançamento, automação e análise de dados. Lucas valoriza a coleta de dados precisos e relevantes, bem como a integração de informações de diversas fontes. Ele está em busca de agilidade e praticidade em suas análises. <p>"Preciso aprimorar minhas estratégias de go-to-market. As atuais não estão fornecendo os resultados desejados."</p> <p>Sentimento: Determinado</p>	<ul style="list-style-type: none"> Ele investe na implementação da plataforma e capacita sua equipe para utilizá-la. <p>"Vamos fazer um investimento na implementação e treinamento para que todos possam aproveitar ao máximo a plataforma."</p> <p>Sentimento: Empolgado</p>	<ul style="list-style-type: none"> Lucas e sua equipe começam a utilizar a solução para coletar, analisar e visualizar dados de marketing e vendas. Eles configuram painéis de controle personalizados que fornecem insights em tempo real. Lucas percebe que a plataforma facilita análises avançadas e tomadas de decisões estratégicas. <p>"Esta solução facilita a análise avançada e torna nossas decisões mais informadas."</p> <p>Sentimento: Confuso</p>	<ul style="list-style-type: none"> Lucas e sua equipe enfrentam desafios técnicos iniciais, como a configuração e integração de fontes de dados complexas. Com o suporte técnico da AWS e treinamento adicional, eles superam esses obstáculos. Lucas aprimora suas habilidades de análise de dados ao longo do tempo. <p>"Estou me tornando mais proficiente na análise de dados à medida que superamos esses obstáculos."</p> <p>Sentimento: Confiante</p>	<ul style="list-style-type: none"> Com o uso contínuo da solução, Lucas e sua equipe veem resultados significativos. Lucas alcança seus objetivos de go-to-market de forma consistente e demonstra um retorno sólido sobre o investimento. <p>Sentimento: Muito Satisfeito e otimista</p>

4.2.2.9 Desfecho

- Após o usuário se habituar com o sistema, é esperado que ele consiga uma melhora significativa na qualidade dos insights obtidos e que aumente velocidade que as análises são feitas, assim como o volume de informações fique mais granular

4.2.2.10 Insights

- O ponto crucial é na Fase 2, onde o usuário está nos primeiros estágios da utilização do sistema e existe a chance dele não se adaptar com a interface e do jeito que os dados se apresentam. Percebe-se que é fundamental que o cliente se adapte ao uso da ferramenta, antes que ele desista de utilizá-la. Por mais que o cliente esteja disposto a aprender a ferramenta, um infográfico bom e intuitivo é indispensável.

4.3 User Stories

A User Story é uma técnica utilizada em desenvolvimento de software para descrever um requisito do ponto de vista do usuário final. Essas histórias são sucintas, centradas no usuário e orientadas para resultados, com o objetivo de comunicar de forma eficaz o que precisa ser construído. Cada user story geralmente inclui um título que descreve o recurso desejado, uma descrição detalhada do comportamento esperado e critérios de aceitação que determinam quando a história está completa. Essa abordagem

permite que um desenvolvimento com foco contínuo nas necessidades dos usuários e forneçam valor de maneira iterativa ao longo do ciclo de desenvolvimento do software.

4.3.1 User Stories - Lucas Macedo

User Story #01
Título: Infográfico Interativo para Tomada de Decisões no Marketing
Persona: Lucas, Consultor de design e marketing
História: <i>Eu, como um consultor de marketing, gostaria de ter um infográfico interativo, a fim de visualizar as informações e métricas sobre o potencial de consumo de cada categoria (canal de atendimento e região) e tomar decisões assertivas na estratégia de lançamento a mercado.</i>
Critérios de avaliação Critério 1: O infográfico é interativo e permite a exploração de dados. Critério 2: O infográfico é atualizado automaticamente com os dados mais recentes.. Critério 3: É possível personalizar o infográfico de acordo com as necessidades do usuário
Testes de aceitação

Critério 1

Aceito: O infográfico permite a interação do usuário, como a capacidade de clicar em elementos para obter detalhes adicionais.

Recusado: O infográfico é estático e não oferece funcionalidades interativas.

Critério 2

Aceito: O infográfico é conectado aos dados em tempo real ou é atualizado automaticamente em intervalos regulares.

Recusado: O infográfico não é atualizado com os dados mais recentes, requerendo atualizações manuais.

Critério 3

Aceito: O usuário pode personalizar o infográfico, escolhendo as métricas e dados a serem exibidos.

Recusado: O infográfico não oferece opções de personalização.

User Story #02

Título

Ferramenta de Análise de Concorrência para Estratégias de Lançamento a Mercado

Persona: Lucas, Consultor de design e marketing

História: Eu, como um consultor de marketing, gostaria de um sistema que auxilie a análise do mercado concorrente a fim de direcionar as estratégias de

lançamento a mercado.

Critérios de avaliação

Critério 1 - As informações coletadas são organizadas e armazenadas de maneira acessível.

Critério 2 - O sistema fornece análises e insights a partir dos dados coletados.

Critério 3 - O sistema é seguro e mantém a confidencialidade das informações coletadas.

Testes de aceitação

Critério 1

Aceito: *As informações sobre os concorrentes são organizadas em um formato acessível e de fácil consulta.*

Recusado: *As informações são desorganizadas ou difíceis de acessar.*

Critério 2

Aceito: *O sistema é capaz de gerar análises e insights úteis a partir dos dados dos concorrentes, capaz de gerar impacto nas decisões estratégicas*

Recusado: *O sistema não gera análises ou insights úteis.*

Critério 3

Aceito: *Medidas de segurança são implementadas para proteger a confidencialidade das informações da empresa e dos concorrentes.*

Recusado: O sistema não é seguro e não protege adequadamente as informações confidenciais.

User Story #03

Título

Sistema de Análise de Vendas para Otimização de Investimentos em Produtos

Persona: Lucas, Consultor de design e marketing

História: Eu, como um consultor de marketing, gostaria de um sistema que auxilie a análise das vendas a fim de definir a quais produtos direcionar mais investimentos.

Critérios de avaliação

Critério 1 - O sistema permite a coleta de dados de vendas de produtos.

Critério 2 - As informações coletadas são organizadas e armazenadas de maneira acessível.

Critério 3 - Permite a comparação de desempenho de diferentes produtos.

Testes de aceitação

Critério 1

Aceito: O sistema é capaz de coletar dados detalhados de vendas de produtos, incluindo informações sobre unidades vendidas, receita gerada e datas de vendas.

Recusado: O sistema não coleta dados de vendas de produtos de forma adequada.

Critério 2

Aceito: As informações de vendas são organizadas em um formato acessível e de fácil consulta, facilitando o trabalho do responsável.

Recusado: As informações são desorganizadas ou difíceis de acessar

Critério 3

Aceito: O sistema permite comparar o desempenho de diferentes produtos em várias métricas de vendas, permitindo a análise.

Recusado: Não é possível comparar o desempenho de produtos de forma eficaz.

User Story #04

Título

Estratégia de lançamento de produto

Persona: Lucas, Consultor de design e marketing

História: Eu como consultor de design e marketing, gostaria de a partir de insights fornecidos, e garantir o lançamento do produto da respectiva área.

Critérios de avaliação

Critério 1 - A ferramenta deve ser capaz de capturar e processar dados (de preferência) relevantes para o usuário.

Critério 2 - A ferramenta deve integrar insights de diferentes fontes, como pesquisas de mercado, análises de concorrência, e exportar relatórios

Critério 3 - Ferramenta precisa ter acesso simples e garantir que o consultor possa acessar os insights de forma fácil e intuitiva através da interface do sistema.

Testes de aceitação

Critério 1

Aceito: *Retorna dados coerentes, e precisos.*

Recusado: *Não retorna nenhum dado, ou dados imprecisos.*

Critério 2

Aceito: *O sistema integra dados de pelo menos três fontes diferentes.*

Recusado: *O sistema não possui interoperabilidade.*

Critério 3

Aceito: *O sistema exibe os resultados filtrados de forma rápida e precisa, permitindo que o consultor acesse facilmente as informações necessárias.*

Recusado: O sistema não se adequa aos filtros selecionados.

4.3.2 User Stories - Ricardo Silva

User Story #05

Título

Configurar o ambiente AWS para o pipeline de Big Data

Persona: Ricardo, cientista de dados

História: Eu, Como analista de dados, desejo que o ambiente AWS seja configurado com as instâncias, serviços e recursos necessários para o pipeline de Big Data, a fim de realizar análises estatísticas em dados armazenados no datalake ou data warehouse.

Critérios de avaliação

Critério 1 - A infraestrutura está pronta para processar e armazenar os dados.

Critério 2 - Os recursos de rede, segurança e permissões estão definidos conforme as melhores práticas.

Critério 3 - Todas as instâncias AWS necessárias estão provisionadas.

Testes de aceitação

Critério 1

Aceito: A infraestrutura está pronta para processar e armazenar os dados

conforme especificado.

Recusado: *Há problemas de desempenho, capacidade insuficiente ou outros obstáculos que impedem o processamento e o armazenamento eficaz dos dados.*

Critério 2

Aceito: *As políticas de segurança e permissões estão configuradas corretamente, seguindo as melhores práticas da AWS.*

Recusado: *Existem lacunas nas políticas de segurança ou permissões que representam riscos de segurança.*

Critério 3

Aceito: *Todas as instâncias estão criadas e em execução*

Recusado: *Pelo menos uma instância não está em execução ou está com problemas de configuração.*

User Story #06

Título

Sistema de Input Flexível para Adaptação a Diferentes Casos de Clientes e
Usuários para Cientistas de Dados

Persona: Ricardo, Cientista de dados

História: *Eu, como cientista de dados, gostaria de um sistema que contém um input flexível para que se adapte a diferentes casos de clientes e usuários.*

Critérios de avaliação

Critério 1 - O sistema suporta configurações personalizadas para diferentes casos de clientes.

Critério 2 - Os usuários podem definir regras de negócios e lógica personalizada.

Critério 3 - O input flexível é mantido de forma segura e confidencial.

Testes de aceitação

Critério 1

Aceito: *O sistema permite a configuração de parâmetros e funcionalidades específicas para atender às necessidades de diferentes clientes.*

Recusado: *O sistema não oferece suporte para configurações personalizadas.*

Critério 2

Aceito: *Os usuários têm a capacidade de definir regras de negócios e lógica personalizada para a análise de dados.*

Recusado: *A definição de regras de negócios e lógica personalizada não é suportada pelo sistema.*

Critério 3

Aceito: *Medidas de segurança são implementadas para proteger a confidencialidade e integridade dos dados de entrada flexíveis.*

Recusado: *O sistema não mantém a segurança adequada dos dados de entrada*

flexíveis.

User Story #07

Título

Integração Flexível com Ferramentas para Facilitar Análises e Insights para Cientistas de Dados

Persona: Ricardo, cientista de dados

História: *Eu, como cientista de dados, gostaria de um sistema que possibilite a integração com diferentes ferramentas a fim de facilitar as análises e a geração de insights.*

Critérios de avaliação

Critério 1 - O sistema suporta integração com uma variedade de ferramentas de análise de dados.

Critério 2 - O sistema permite a transferência eficiente de dados entre as ferramentas integradas.

Critério 3 - O sistema suporta a automatização de fluxos de trabalho com as ferramentas integradas.

Testes de aceitação

Critério 1

Aceito: *O sistema pode ser integrado a uma ampla gama de ferramentas de análise de dados, como Python, R, Tableau, Power BI, e outras.*

Recusado: O sistema oferece suporte limitado ou nenhuma integração com ferramentas de análise de dados.

Critério 2

Aceito: A transferência de dados entre o sistema e as ferramentas integradas é rápida e eficiente. Sem a necessidade de muito trabalho manual.

Recusado: A transferência de dados é lenta ou ineficiente.

Critério 3

Aceito: Os fluxos de trabalho podem ser automatizados com as ferramentas integradas, economizando tempo e recursos.

Recusado: A automatização de fluxos de trabalho é complexa ou não é suportada.

User Story #08

Título

Sistema com Filtros Versáteis para Análises Dinâmicas de Dados

Persona: Ricardo, cientista de dados

História: Eu, como cientista de dados, gostaria de um sistema que ofereça filtros a fim de realizar diferentes análises (descritiva, diagnóstica e entre outros).

Critérios de avaliação

Critério 1 - O sistema oferece filtros por região, clientes e vendas.

Critério 2 - O sistema oferece filtros de potenciais clientes e potenciais mercados.

Critério 3 - O sistema apresenta os filtros de maneira clara, coesa e funcional.

Testes de aceitação

Critério 1

Aceito: Os filtros por região, clientes e vendas estão disponíveis no sistema. Eles funcionam corretamente e permitem ao usuário filtrar os dados com sucesso.

Recusado: Os filtros por região, clientes e vendas não estão disponíveis no sistema. Eles não funcionam como esperado ou estão ausentes.

Critério 2

Aceito: O sistema oferece filtros de potenciais clientes e potenciais mercados. Eles são funcionais e permitem ao usuário filtrar com sucesso as informações relacionadas a esses critérios.

Recusado: O sistema não oferece filtros de potenciais clientes e potenciais mercados, ou esses filtros não funcionam adequadamente.

Critério 3

Aceito: Os filtros são apresentados de forma clara e coesa no sistema.

Recusado: Os filtros não são apresentados de maneira clara e coesa. Eles são confusos, difíceis de usar ou não funcionam.

5. Arquitetura Macro

5.1 Tipos de dados fornecidos pelo parceiro e suas características.

5.1.1 Formato dos Dados de Entrada Os dados vêm de forma não estruturados, com diferentes formatos de arquivos, desde CSV até arquivos em texto.

5.1.2 Fontes de Dados:

- Dados de censos/pesquisas do governo (CSV)
- Dados de CNPJs (CSV)
- Dados fornecidos pelo parceiro (API)

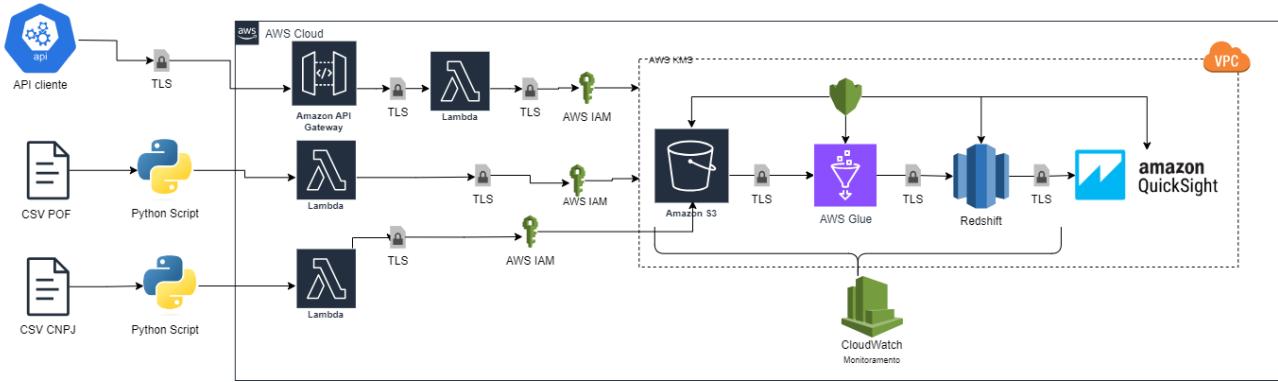
5.2 Requisitos do pipeline de dados:

- **Volume de Dados:**
 - Há variação no volume de dados devido às contribuições do parceiro e ao crescimento contínuo ao longo dos anos nos dados governamentais e de CNPJs
 - Planejamento para suportar um volume de dados médio até 10GB
- **Velocidade de Ingestão:**

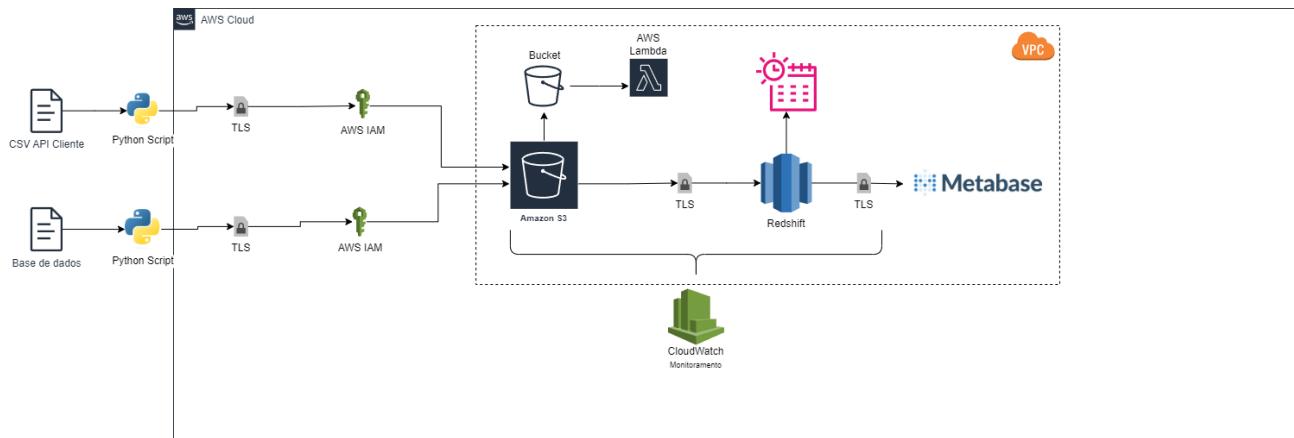
- A Transmissão dos dados será via streaming
 - Processamento dos fluxos durante a visualização das informações do infográfico
- **Transformação e Processamento:**
 - Dados recebidos em formato não estruturado
 - Transformação em tabelas estruturadas durante o processamento (principalmente no processo de ETL)
 - Limpeza, remoção de dados indesejáveis (de início apenas os dados estrangeiros foram salientados como necessidade de remoção), reestruturação do formato dos dados
- **Armazenamento:**
 - Armazenamento dos dados a partir de Amazon S3 e do AWS Redshift
 - Há possibilidade de portabilidade para outras plataformas de nuvem, usando serviços de código aberto
- **Segurança:**
 - Utilização do Amazon TLS
 - Utilização do AWS VPC que fornece um ambiente de rede virtual seguro e isolado
- **Escalabilidade:**
 - Gerenciamento eficaz de requisições mesmo com grandes volumes de dados, graças à utilização de serviços que permitem e garantem o desempenho a partir das regras de negócio.

5.3. Arquitetura

5.3.1 Arquitetura Inicial



5.3.2 Arquitetura Final Completa



5.4 Descrição do Fluxo de Dados da Arquitetura

O processo começa com o API Gateway recebendo as solicitações de entrada, a partir dos dados do cliente provisionados por uma API própria. Além disso, possuímos outras duas fontes de entrada, dados de pesquisas do governo e informações de CNPJs, que são coletadas, tratadas e processadas inicialmente por Scripts em Python, provisionados posteriormente em Lambda Functions para envio dos dados ao datalake.

O datalake de armazenamento utilizado é o Amazon S3, que disponibiliza e provém escalabilidade dos dados para qualquer tipo de consulta ou requisição de serviço, no qual

os recursos de gerenciamento são fáceis de usar, e há uma maior organização dos dados, além da possibilidade de configurar controles de acesso para atender a requisitos específicos de negócios.

Para permitir a comunicação segura entre as Lambda Functions e o Amazon S3, temos o AWS Identity and Access Management (IAM). O AWS IAM é o serviço de gerenciamento de identidade e acesso da AWS que permite controlar quem pode acessar seus recursos da AWS e o que eles podem fazer com esses recursos.

Em seguida, o AWS Lambda que já provisionou os códigos permite sua requisição em tempo real e para disponibilidade no S3. Assim, a partir da extração destes dados, o AWS Glue desempenha um papel crucial nessa fase, organizando os dados brutos ao realizar a limpeza e as transformações necessárias (processo de ETL). Posteriormente, ele prepara esses dados para armazenamento.

Após a transformação, os dados são encaminhados para o armazenamento permanente. Eles são armazenados no Amazon Redshift. Para garantir a segurança durante a ingestão, são estabelecidas várias medidas de segurança. A aplicação da comunicação de conexões seguras (TLS) são utilizadas entre todos os componentes do processo. Além disso, políticas rigorosas de controle de acesso são configuradas no Amazon S3 para restringir quem pode acessar os dados. A criptografia é aplicada aos dados usando o AWS KMS, garantindo que permaneçam confidenciais e seguros.

Para monitorar todo o processo, são configurados alarmes no Amazon CloudWatch. Esses alarmes são essenciais para detectar problemas ou gargalos no fluxo de dados, permitindo uma resposta rápida e eficiente para manter o sistema funcionando sem interrupções. Esse fluxo de dados bem organizado e altamente seguro garante que as

informações sejam processadas, transformadas, armazenadas e monitoradas de maneira eficaz e confiável.

5.6 Descrição da funcionalidade dos serviços na arquitetura:

Segue descrição dos serviços e sua funcionalidade, contidos na arquitetura:

1. **API Gateway:** ponto de entrada os dados vindo de API do cliente. Ele lida com as solicitações recebidas e as encaminha para os serviços apropriados.
2. **API Gateway:** ponto de entrada os dados, que projetará as informações vindas de API do cliente. Ele lidará com as solicitações recebidas e fará o redirecionamento pro datalake (Amazon S3). Em suma, o API Gateway permite que os desenvolvedores criem, publiquem, mantenham, monitorem e protejam APIs em qualquer escala, o que auxilia na escalabilidade do projeto.
3. **Lambda Function:** Uma vez que o API Gateway recebe a solicitação, ele aciona uma função Lambda, e no outro lado de recebimento dos dados, os Scripts gerados em python também serão utilizados no Lambda. O Lambda é um serviço de computação sem servidor que executa seu código em resposta a eventos e gerencia automaticamente os recursos de computação em nuvem.
4. **Amazon S3 Bucket:** A partir disso, as funções Lambda processam as solicitações e armazena os dados em um bucket Amazon S3. O Amazon S3 é um serviço de armazenamento de objetos que oferece alta escalabilidade e é utilizada na arquitetura como datalake para armazenamento, disponibilidade de dados, com alta segurança (a partir das pesquisas realizadas) e desempenho.
5. **Amazon Glue:** Na seção de transformação dos dados, selecionamos o AWS Glue, que receberá os dados do S3. O Amazon Glue é um serviço de ETL totalmente

gerenciado (permite alto controle do serviço) que facilita a preparação e o carregamento de dados para análise.

6. **Amazon Redshift Cluster:** Os dados do bucket S3 são então movidos para o AWS Glue e depois desta parte de ETL, serão movidos para um cluster do Amazon Redshift para processamento adicional. O Amazon Redshift é um serviço de armazenamento de dados em nuvem totalmente gerenciado que permite executar consultas SQL de dados, e neste caso será utilizado como nosso datawarehouse
 1. É importante citar que o Redshift roda com Spark “por baixo dos panos” e possui alta eficiência para dados de alto volume e complexidade
7. **Amazon CloudWatch:** Durante todo esse processo, o Amazon CloudWatch monitora os recursos, desde o processo no datalake até o processo de datawarehouse, e coleta e rastreia métricas, arquivos de log e responde a alterações de desempenho em todo o sistema.
8. **Amazon QuickSight:** Finalmente, os dados processados do cluster Redshift são visualizados usando o Amazon QuickSight, um serviço de inteligência empresarial (BI) escalável, sem servidor, incorporável e alimentado por machine learning construído para a nuvem.
9. **Amazon VPC:** Todos esses componentes estão contidos dentro de uma Amazon Virtual Private Cloud (VPC), que fornece um ambiente de rede virtual seguro e isolado.
10. **AWS Identity and Access Management (IAM):** pode ser definido como um serviço da Amazon Web Services (AWS) que permite controlar o acesso aos recursos da AWS de forma segura e escalável. A funcionalidade do AWS IAM é essencial para gerenciar identidades e permissões dentro do ambiente, e neste

caso, permitirá o acesso granular a cada uma das aplicações na AWS de nossa arquitetura.

5.7 Canais e Métodos de Ingestão

5.7.1 API Gateway:

O API Gateway serve como o ponto de entrada do sistema, recebendo solicitações de entrada da API do cliente. Ele é especialmente adequado para lidar com dados em tempo real, onde a frequência é alta e as solicitações são recebidas de maneira contínua. Este canal é projetado para fornecer escalabilidade e segurança para gerenciar o fluxo constante de dados em tempo real.

Frequência de Dados: Alta, em tempo real.

Quantidade de Dados: Variável, dependendo da interação com os clientes.

Vantagens: Alta escalabilidade, segurança integrada, manipulação de dados em tempo real.

5.7.2 AWS Lambda:

O AWS Lambda é um serviço altamente flexível que desempenha um papel crucial no processamento de dados estáticos e em tempo real. É especialmente valioso quando se lida com dados estáticos, como POF (Pessoa Física) e CNPJ (Cadastro Nacional de Pessoa Jurídica), que têm uma frequência menor e podem ser importados e processados em intervalos programados.

Frequência de Dados Estáticos: Baixa a moderada, em intervalos programados.

Quantidade de Dados Estáticos: Variável, dependendo do volume de dados programados.

Frequência de Dados em Tempo Real: Alta, contínua.

Quantidade de Dados em Tempo Real: Variável, dependendo da interação com os clientes.

Vantagens: Flexibilidade para processar dados estáticos e em tempo real, controle sobre agendamento e escalabilidade.

5.8 Seleção dos Serviços da AWS para Cada Etapa do Processo de Ingestão

5.8.1 Coleta e Ingestão Inicial:

O API Gateway recebe as solicitações de entrada:

- O API Gateway atua como o ponto de entrada para o sistema, recebendo solicitações de entrada da API do cliente. As solicitações são encaminhadas para o próximo estágio do pipeline de dados.

Processamento com AWS Lambda para Dados Estáticos e em Tempo Real:

- Dados Estáticos: Para dados estáticos (POF e CNPJ), o AWS Lambda pode ser configurado para realizar tarefas programadas, como a importação e processamento de arquivos em intervalos específicos. O Lambda processa esses dados de acordo com as necessidades do negócio, aplicando transformações, validações e preparando-os para armazenamento.

- Dados em Tempo Real: Para dados em tempo real (API do Cliente), o AWS Lambda é configurado para processar as solicitações conforme elas chegam. Ele aplica transformações e operações em tempo real aos dados em movimento, garantindo que eles estejam prontos para armazenamento imediato.

Transformação e Limpeza:

AWS Glue para Transformação e Limpeza:

- O AWS Glue desempenha um papel fundamental na etapa de transformação e limpeza dos dados. Ele pode extrair os dados do Amazon S3, aplicar transformações para padronizar, limpar e enriquecer os dados, e então armazená-los diretamente no Amazon Redshift.

Armazenamento:

Amazon S3 como Datalake:

- O Amazon S3 é usado como um data lake para receber os dados brutos. Ele fornece um armazenamento escalável e econômico para armazenar os dados em sua forma bruta, sem estruturação.

Amazon Redshift como Datawarehouse:

- O Amazon Redshift é usado para armazenar os dados preparados e estruturados de forma apropriada para consulta analítica. Os dados processados pelo AWS Glue são carregados no Amazon Redshift para permitir consultas de alto desempenho e análises complexas.

5.9 Aspectos de segurança da Arquitetura

Visando proteger os dados durante o processo de ingestão, utilizaremos recursos específicos: o TLS (Transport Layer Security) e o AWS Key Management Service (KMS), que desempenham papéis na garantia da segurança da comunicação.

O TLS desempenhará um papel no que diz respeito à criptografia. Sua função principal é criptografar os dados enquanto estão em trânsito, o que impede possíveis tentativas de interceptação e assegura a integridade dos dados em todo o percurso, desde a etapa do AWS Glue até o Redshift, por exemplo. Além da criptografia, o TLS também verifica a integridade dos dados, garantindo que nenhuma alteração ocorra durante a transferência.

Para complementar a segurança, vamos fazer uso do AWS Key Management Service (KMS). O KMS pode ser integrado aos serviços que planejamos utilizar e nos permite criar Customer Master Keys (CMKs), que são essenciais para criptografar e decifrar os dados de forma segura. Além disso, o KMS é responsável pela gestão eficiente dessas chaves (CMK). Portanto, o KMS contribui significativamente para a segurança do processo de ingestão de dados, garantindo que as chaves criptográficas utilizadas pelo TLS sejam gerenciadas e protegidas adequadamente. Juntos, o TLS e o KMS se complementam garantindo a integridade e a confidencialidade dos dados durante o trânsito.

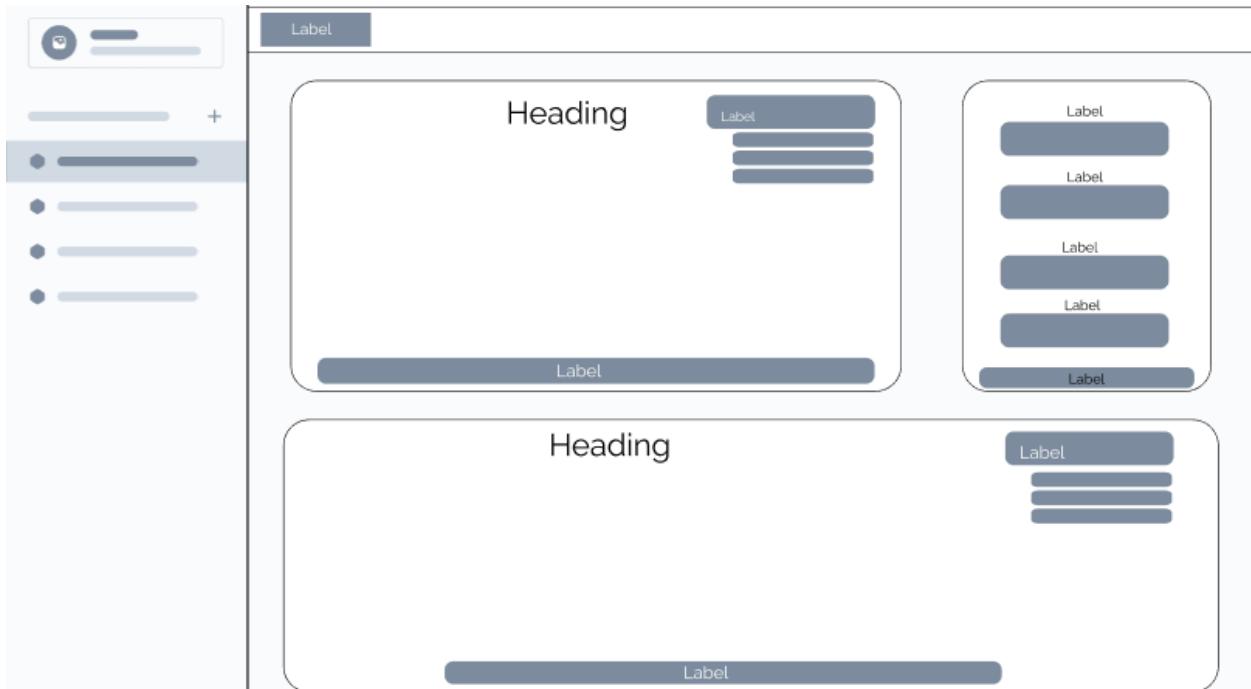
Além disso, adotamos o AWS Identity and Access Management (IAM) para aprimorar a segurança em nossa arquitetura. Essa abordagem nos capacita a estabelecer um controle granular sobre as permissões de acesso dos usuários em suas funções específicas, garantindo um ambiente altamente seguro.

6. Wireframe

O processo de prototipação de baixa fidelidade desempenha um papel importante no processo de design e entendimento da disposição das informações, oferecendo uma abordagem

flexível e eficiente para a concepção inicial de interfaces e futuro embasamento para mockups.

Neste caso, a intenção principal se refere à criação de representações esquemáticas (neste caso simplificadas) da interface do usuário, com o cuidado de não se aprofundar em detalhes visuais ou estilísticos. Segue protótipo inicial do wireframe:



6.1 Justificativa das escolhas de design

As escolhas de design são fundamentais para a criação de uma interface de usuário intuitiva e eficaz. Essas escolhas são informadas por uma variedade de considerações, incluindo clareza, confiança do usuário, simplicidade, visibilidade do estado do sistema e liberdade de controle do usuário. As decisões de design foram moldadas a partir de feedbacks das partes interessadas e terceiros, a fim de melhorar a experiência do usuário e permitir uma navegação ininterrupta e maleável.

Sendo assim, os principais pontos pensados para a criação do wireframe foram:

- Manter a clareza: O wireframe usa rótulos simples e claros para indicar as funções e os conteúdos de cada seção. Os gráficos mostram claramente as informações que estão

mostrando para que não existam dúvidas. Cada gráfico tem uma legenda própria e estão devidamente rotulados.

- Padrões: O wireframe segue padrões de design de interface do usuário que são familiares e intuitivos para os usuários. O menu está no lado esquerdo, o título está no topo e a legenda está na parte inferior dos gráficos. Os botões têm formas e cores consistentes e contrastantes com o fundo.
- Simplicidade é a chave: O wireframe usa formas geométricas, cores sólidas, em conjunto com textos para representar os elementos visuais. O wireframe se concentra na funcionalidade, não desprezando o estilo.
- Visibilidade do estado do sistema: O wireframe foi pensado para que o usuário nunca fique perdido e sempre tenha consciência do que está fazendo/visualizando e onde se encontra.
- Liberdade de controle fácil para o usuário: A partir de feedbacks, concluímos que uma das partes mais importantes do infográfico é a flexibilidade e adaptação das informações de acordo com a necessidade do usuário, então, elementos como filtros são utilizados para compreender essa tarefa e permitir uma navegação interativa com o usuário

6.2 Técnicas ou estratégias avançadas

O wireframe que foi proposto pelo grupo, tomando como partida as informações coletadas nas conversas com o cliente (e especialmente diante do lean inception realizado), contém elementos direcionados para um infográfico (na intenção de uma temática de painel de controle). O wireframe antecipa a necessidade do usuário de filtrar e classificar algumas informações específicas, fornecendo opções de filtro e um menu suspenso. Além disso, o wireframe também inclui micro-interações, como estados de foco para botões e rótulos (focando apenas na disposição e funcionalidade do elemento). Para contextualização, as micro-interações são basicamente pequenos feedbacks que ocorrem quando o usuário interage com a interface. Consequentemente, elas são usadas para fornecer feedback imediato ao usuário, tornando a

experiência mais envolvente e agradável. Especificamente, o botão de filtro e o menu suspenso permitem que o usuário filtre e classifique os dados (principalmente as informações de canal/categoria/região, métricas de vendas, cliente, potencial mercado e cliente, e outras KPI's) com facilidade, sem precisar procurar essas opções em outros lugares. Isso economiza tempo e torna a experiência do usuário mais agradável. Já os estados de foco ajudam o usuário a entender qual elemento está selecionado e o que acontecerá quando ele clicar nele.

No escopo macro, visualizamos espaço para disposição de elementos que são necessários para visualização mais interativa dessas informações. O mapa e o gráfico não estão incluídos diretamente no wireframe, no entanto, são exemplos de como imaginamos a visualização dos dados pode ser apresentada de maneira clara e fácil compreensão. O planejamento inicial é que o mapa mostre as informações sobre as regiões e o número de clientes em cada uma delas, tendo em vista a necessidade do parceiro. Já, na utilização dos gráficos, pensamos em três vertentes específicas que eles seriam relevantes, sendo elas: informações de vendas, cruzamento de vendas com outras informações (exemplo: canal, categoria, região) e potencial de mercado e cliente. Neste contexto, para informações de vendas mostraremos algumas informações sobre as vendas ao longo do tempo, com o pensamento de possíveis utilizações de gráficos de linhas e de área. Na vertente de cruzamento de informações, pensamos em utilizar o gráfico de dispersão (que é bom para fazer este tipo de relação entre variáveis). Por fim, para potencial de mercado e cliente, pensamos em colocar estas informações utilizando o gráfico de barras (de preferência horizontal, se houver espaço para disposição) e o gráfico de funil. Sendo assim, esses elementos ajudam o usuário a entender rapidamente o fluxo de desempenho, potencial de mercado, para que assim, possa tomar decisões informadas.

6.3 Feedback e iterações

Visando melhorar exponencialmente a qualidade do infográfico, o wireframe foi criado levando em consideração não apenas a opinião do grupo, como também do usuário e de terceiros. Através da pesquisa feita pelo grupo alguns pontos foram levados em consideração para a escolha de design, eles são:

- Ênfase nas informações principais: Foi constatado que um ponto essencial é dispor os elementos considerando a importância para o usuário de cada informação. Garantir que KPI's importantes sejam posicionadas de maneira central no infográfico e que gráficos que apresentam informações mais urgentes ocupem um espaço maior que outros são formas que foram identificadas como úteis a fim de garantir esta priorização.
 - Menu suspenso a fim de tornar claras as funcionalidades disponíveis;
 - Gráfico sobre potencial cliente e mercado feito horizontalmente.
- Tipos de gráficos adequados para cada caso: É essencial que os gráficos utilizados se adequem às finalidades específicas de cada caso. Como citado anteriormente, os gráficos de linhas e de área, por exemplo, podem ser utilizados para informações de venda, enquanto o de funil e o de barras podem ser utilizados para gráficos sobre potenciais mercados e clientes. Os gráficos devem garantir uma boa visibilidade para as informações apresentadas.
 - Gráficos de funil e de barras para potencial mercado e cliente;
 - Gráficos de linha e de área para informações de venda;
 - Gráfico de dispersão para cruzamento de informações.
- Filtros que atendam às necessidades: Os filtros devem permitir a visualização de informações como canal, categoria e região, dados sobre os clientes, potencial mercado e cliente, além de outras KPI's. É importante que os filtros sejam dispostos de forma que denote suas funcionalidades a fim de evitar possíveis confusões por parte dos usuários que ainda não estiverem acostumados com o uso do infográfico.
 - Informações de canal/categoria/região;
 - Métricas de vendas;
 - Dados sobre o cliente;
 - Potencial mercado e cliente;
 - Outras KPI's.

6.3.1 Interação direta com o parceiro

Após a exposição dos wireframes, recebemos feedback tanto do parceiro quanto do professor, os quais estão sendo integrados à evolução dos protótipos e serão apresentados nas próximas entregas. As observações recebidas incluem:

- Recomendação para sempre considerar a análise cruzada de dados de categoria e canal, proporcionando ao usuário uma visão mais abrangente das informações;
- Sugestão de incorporar um filtro de data, permitindo que o usuário selecione o período temporal desejado para a visualização;
- Dimensão geográfica: a sugestão foi de que, quanto mais detalhada, melhor, considerando que o usuário pode ter interesse em analisar dados de uma região específica;
- No painel, a orientação é incluir uma combinação equilibrada de gráficos, tabelas e KPIs. Sugere-se a substituição do ranking por alguns KPIs para melhorar a apresentação visual.

6.3.2 Embasamento nas heurísticas

WIREFRAME

Heurísticas base para desenvolvimento:

- Liberdade de controle do usuário
- Visibilidade do estado no sistema
- Consistência e padrões
- Flexibilidade e eficiência de uso



Como visto anteriormente na seção de “Justificativa das escolhas de design”, alguns pontos foram fortemente influenciados pelas heurísticas de Nielsen. As principais utilizadas foram:

- Consistência e padrões: Prezamos pela padronização dos gráficos para que a navegação fique dinâmica, permitindo que o usuário consiga manipular diversos gráficos de maneiras semelhantes. Esse ponto também é importante para que padrões que o usuário está acostumado estejam da maneira que ele espera, como por exemplo, a barra de menu na parte esquerda da tela
- Visibilidade do estado do sistema: É de extrema importância que o usuário saiba exatamente o que está olhando e que parte do sistema ele se encontra. Então, como é visível pela parte destacada no menu, permite-se a visualização de que grupo de gráficos está sendo visto.
- Liberdade de controle do usuário: Por meio de filtros e ferramentas da plataforma escolhida, o usuário poderá livremente escolher exatamente o que deseja visualizar dentro de um visual, e que informações são importantes para a situação que ele deseja analisar, possibilitando um controle livre.
- Flexibilidade e eficiência de uso: como dito no “Liberdade de controle do usuário”, é essencial que seja possível que o usuário visualize o que deseja, então os filtros entram em ação. O sistema será carregado com um grande gama de informações desejadas pelo cliente e possibilitará as que ele deseja visualizar. O usuário terá uma alta capacidade de capacidade de configurar o que ele deseja ver, aumentando a flexibilidade e eficiência do uso.

7. Estruturação do Datalake

7.1 Identificação das Fontes de Dados e suas características

Neste contexto, é essencial compreender a natureza das diferentes bases de dados, classificando-as entre públicas e privadas, a fim de otimizar seu uso e garantir conformidade com regulamentos de privacidade.

7.1.1 Fontes Principais:

Dados Públicos:

1. Base POF - 500 MB (CSV, XLS)

- Utilização de 11 tabelas
- Frequência de utilização: de acordo com o conjunto de dados que utilizaremos, a sua frequência de atualização é anual (*aproximadamente*)

2. Base IBGE - 3 GB (CSV)

- A princípio, com utilização de duas tabelas
 - Prioridade para *Índice Nacional de Preços ao Consumidor; O PIB por município.*
- Frequência de utilização: anual

3. Base CNPJ - 4,6 GB (CSV)

- Utilização de 5 tabelas
- Frequência de utilização: frequência será anual, tendo em consideração o tempo de pesquisa atribuído

4. Base Dados Geográficos Brasileiros - Tamanho Não identificado (base de dados não operante - CSV)

- Tabelas ainda não utilizadas
- Importância que deve ser considerada para sua utilização: Os dados geográficos brasileiros foram selecionados para entender as demandas regionais.
- Frequência de atualização: trimestralmente

5. Base de Dados Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD-C) - CSV

- Tabelas ainda não utilizadas
- Importância que deve ser considerada para sua utilização: inferência da situação social das famílias brasileiras
- Frequência de atualização: não listada

Dados Privados:

- A princípio serão fornecidos dados fictícios simulando a base de vendas de um distribuidor (conforme categorias, produtos, canais de venda, data) em tempo real, com base nas informações que o parceiro possuem. Serão disponibilizados por meio da API disponibilizada. Estes dados nos darão um embasamento maior e agregaram às outras fontes de dados.

7.2 Seleção dos serviços da AWS

A escolha dos serviços desempenha um papel crucial na construção de uma infraestrutura de armazenamento eficiente e segura na nuvem. Ao abordar a necessidade de armazenamento de dados e a criação de um datalake inicial, a opção pelo Amazon S3 destaca-se como a melhor escolha em termos estratégicos.

Ao optar pelo S3 para o armazenamento na nuvem, há maior capacidade de dimensionar verticalmente conforme as demandas crescentes de dados (tendo em vista que estaremos utilizando diversas fontes de dados de entrada). O serviço oferece uma solução confiável e durável para armazenamento de objetos, garantindo a integridade e disponibilidade dos dados. Além disso, a simplicidade na configuração e administração do Amazon S3 torna o processo de gerenciamento de dados mais eficiente e acessível.

7.2.1 Segurança de dados:

A segurança é outro aspecto fundamental para todo o fluxo que o grupo evidência, e o S3 oferece recursos avançados de controle de acesso e criptografia para proteger os dados armazenados, que podem ser “setados” nas configurações iniciais de cada bucket

- Amazon Key Management System (KMS) → Há possibilidade de implementação do serviço KMS nos buckets de armazenamento. Neste caso, como estrutura acadêmica não implementamos a partir da limitação do uso dos serviços no Labs. No entanto é um aspecto de vital importância na construção de uma infraestrutura segura e eficiente. O KMS desempenha um papel crucial na gestão de chaves de criptografia, fornecendo uma camada adicional de segurança para os dados armazenados na nuvem.
- Virtual Private Cloud (VPC) → A VPC oferece a capacidade de criar segmentos na nuvem da AWS, permitindo o controle de acesso aos serviços utilizados. Isso inclui a implementação de restrições de acesso e regras de segurança, proporcionando uma camada adicional de proteção à solução.

7.3 Processo de Ingestão de Dados e Mapa Mental dos Buckets

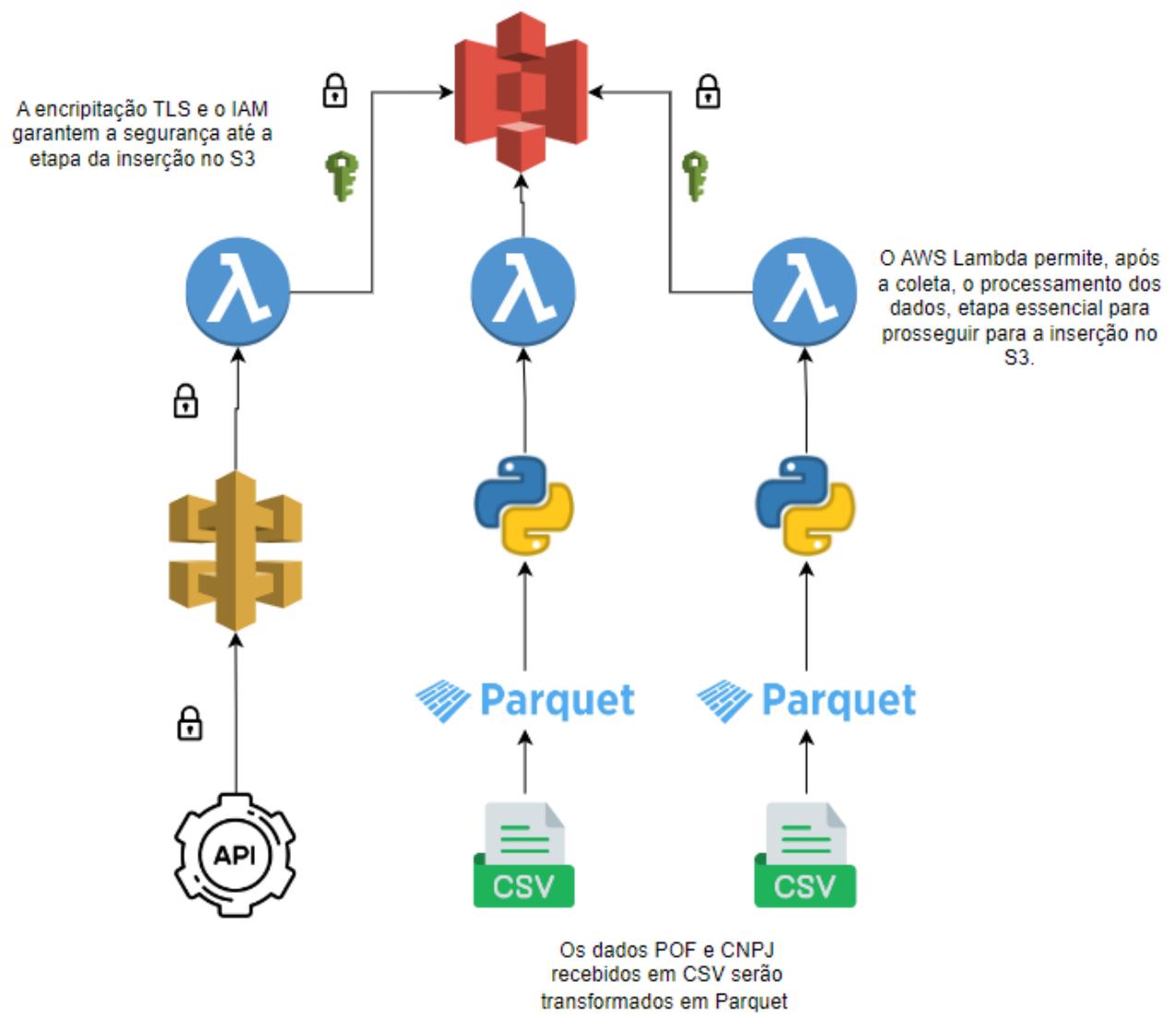
Nesta etapa da arquitetura, a Ingestão de Dados, compreende duas abordagens principais:

- O carregamento dos dados em formato Parquet (convertidos de CSV para Parquet) para a infraestrutura na nuvem.
 - Mais especificamente para Buckets no S3
- A transferência em tempo real dos dados do cliente através da API para o ambiente na nuvem.

Agora, abordaremos detalhadamente o primeiro método de ingestão:

Utilizando scripts em Python, efetua-se o envio de arquivos Parquet para buckets específicos na Amazon S3.

- Cada fonte de dados possui seu próprio bucket designado
- A representação visual abaixo oferece uma visão organizacional clara do processo de ingestão de dados:



7.4 Estrutura do Datalake

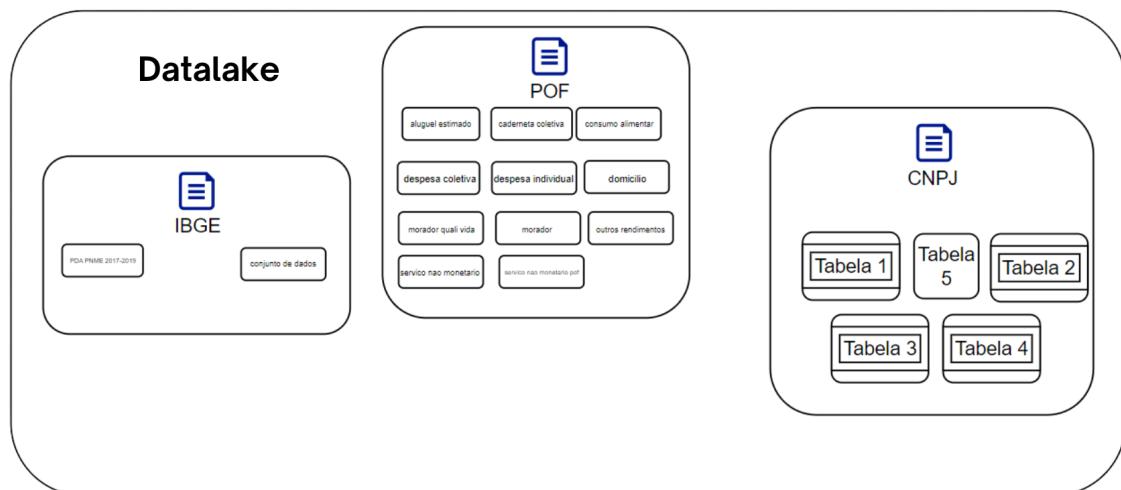
Para contextualização, o Datalake é basicamente um repositório que armazena grandes volumes de dados brutos, estruturados e não estruturados, em seu formato original. Diferente de um data warehouse, que segue um esquema de dados estruturado, o datalake permite realizar o armazenamento dos dados em sua forma bruta, sem a necessidade de uma estrutura pré-definida.

Isso significa que os dados podem ser coletados e armazenados de diferentes fontes, neste caso das formas especificadas. No contexto do projeto, tendo em vista o recebimento de diferentes fontes de dados, isto nos permite apenas se preocupar com etapas iniciais de pré-processamento. A princípio os dados são enviados em formato CSV e Parquet.

Neste caso, o serviço do datalake selecionado foi o S3 (com buckets específicos para cada uma das fontes).

Segue mapa mental, mais detalhista, do uso das fontes de dados e suas respectivas tabelas dentro do datalake:

MAPA MENTAL DATALAKE



7.5 Dados carregados no Datalake (Amazon S3)

7.5.1 Configuração do Bucket:

Passo 1: Acesse o Console da AWS:

- Vá para a página inicial da AWS (<https://aws.amazon.com/>), faça login na sua conta e acesse o Console da AWS.

Passo 2: Navegue até o Amazon S3:

- No Console da AWS, encontre o serviço "S3" ou digite "S3" na barra de pesquisa e selecione-o.

Passo 3: Inicie o Processo de Criação:

- Dentro do painel do Amazon S3, clique no botão "Criar bucket".

Passo 4: Configure as Opções do Bucket:

- Preencha as informações necessárias, incluindo um nome **único** para o seu bucket. O nome do bucket deve ser globalmente exclusivo, já que é usado no URL do bucket.
- Escolha, também, a região onde o bucket será armazenado.

Passo 5: Configure as Propriedades do Bucket:

- Configure as opções adicionais, como versão do bucket, logging, permissões de bloqueio, entre outros.

Criar bucket Informações

Os buckets são contêineres para dados armazenados no S3. [Saiba mais](#)

Configuração geral

Nome do bucket

O nome do bucket deve ser exclusivo no namespace global e seguir as regras de nomenclatura do bucket. [Veja as regras para nomenclatura de buckets](#)

Região da AWS



Copiar configurações do bucket existente - *opcional*

Somente as configurações de bucket na configuração a seguir são copiadas.

Propriedade de objeto Informações

Controle a propriedade de objetos gravados nesse bucket a partir de outras contas da AWS e o uso de listas de controle de acesso (ACLs). A propriedade do objeto determina quem pode especificar o acesso aos objetos.

ACLs desabilitadas (recomendado)

Todos os objetos nesse bucket são de propriedade dessa conta. O acesso a esse bucket e seus objetos é especificado usando apenas políticas.

ACLs habilitadas

Os objetos nesse bucket podem ser de propriedade de outras contas da AWS. O acesso a esse bucket e seus objetos pode ser especificado usando ACLs.

Propriedade do objeto

Imposto pelo proprietário do bucket

Passo 6: Configure as Permissões do Bucket:

- Defina as permissões do bucket, especificando quem pode acessar e modificar o conteúdo do bucket. Você pode configurar políticas de controle de acesso aqui.

Configurações de bloqueio do acesso público deste bucket

O acesso público é concedido a buckets e objetos por meio de listas de controle de acesso (ACLs), políticas de bucket, políticas de ponto de acesso ou todas elas. Para garantir que o acesso público a este bucket e todos os seus objetos seja bloqueado, ative a opção de Bloquear todo o acesso público. Essas configurações serão aplicadas apenas a este bucket e aos respectivos pontos de acesso. A AWS recomenda ativar a opção Bloquear todo o acesso público. Porém, antes de aplicar qualquer uma dessas configurações, verifique se as aplicações funcionarão corretamente sem acesso público. Caso precise de algum nível de acesso público a este bucket ou aos objetos que ele contém, é possível personalizar as configurações individuais abaixo para que atendam aos seus casos de uso de armazenamento específicos. [Saiba mais](#)

Bloquear todo o acesso público

Ativar essa configuração é o mesmo que ativar todas as quatro configurações abaixo. Cada uma das configurações a seguir são independentes uma da outra.

Bloquear acesso público a buckets e objetos concedidos por meio de *novas* listas de controle de acesso (ACLs)

O S3 bloqueará as permissões de acesso público aplicadas a blocos ou objetos recém-adicionados e impedirá a criação de novas ACLs de acesso público para blocos e objetos existentes. Essa configuração não altera nenhuma permissão existente que permita o acesso público aos recursos do S3 usando ACLs.

Bloquear acesso público a buckets e objetos concedidos por meio de *qualquer* lista de controle de acesso (ACLs)

O S3 ignorará todas as ACLs que concedem acesso público a buckets e objetos.

Bloquear acesso público a buckets e objetos concedidos por meio de *novas* políticas de ponto de acesso e bucket público

O S3 bloqueará novas políticas de bucket e ponto de acesso que concedem acesso público a buckets e objetos. Essa configuração não altera nenhuma política existente que permita o acesso público aos recursos do S3.

Bloquear acesso público e entre contas a buckets e objetos por meio de *qualquer* política de bucket ou ponto de acesso público

O S3 ignorará o acesso público e entre contas para buckets ou pontos de acesso com políticas que concedem acesso público a buckets e objetos.

Passo 7: Revise as Configurações e Crie o Bucket:

- Revise todas as configurações feitas para garantir que estejam corretas. Clique em "Criar bucket" para confirmar e criar o novo bucket.

Criar bucket

Passo 8: Acesse o Bucket:

- Após a criação, você verá seu novo bucket listado no painel do Amazon S3. Clique nele para acessar e começar a gerenciar objetos (arquivos) dentro do bucket.

7.6 Dados alimentados no Datalake (exemplo CNPJ e POF)

Separação dos buckets com base em cada uma das fontes de dados:

The screenshot shows the Amazon S3 Buckets page. At the top, there's a summary section with metrics like 'Snapshot da conta' (Storage Lens), 'Armazenamento total' (Pending), 'Quantidade de objetos' (Pending), and 'Tamanho médio do objeto' (Pending). Below this is a table titled 'Buckets (2) Informações' showing two buckets:

Nome	Região da AWS	Acesso	Data de criação
grupo-5-mec-solutions-bucket-dados-cnpj	Leste dos EUA (Norte da Virgínia) us-east-1	Bucket e objetos não públicos	9 Nov 2023 05:04:42 PM -03
grupo-5-mec-solutions-bucket-dados-pof	Leste dos EUA (Norte da Virgínia) us-east-1	Bucket e objetos não públicos	12 Nov 2023 10:54:56 PM -03

Tabelas *tratadas e que subidas automaticamente no datalake:

The screenshot shows the Amazon S3 Object details page for the 'grupo-5-mec-solutions-bucket-dados-pof' bucket. The top navigation bar includes links for 'Bucket', 'Informações', 'Objetos', 'Propriedades', 'Permissões', 'Métricas', 'Gerenciamento', and 'Pontos de acesso'. The main area displays a table of objects:

Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
aluguel_estimado.csv	csv	12 Nov 2023 11:40:51 PM -03	5.5 MB	Padrão
domicilio_processado.csv	csv	12 Nov 2023 11:43:00 PM -03	9.9 MB	Padrão
outros_rendimentos.csv	csv	13 Nov 2023 12:03:02 AM -03	32.8 MB	Padrão
servico_nao_monetario_pof2.csv	csv	12 Nov 2023 11:54:01 PM -03	2.8 MB	Padrão
servico_nao_monetario_pof4.csv	csv	12 Nov 2023 11:47:30 PM -03	21.5 MB	Padrão

*Todo o processo de envio dos dados está sendo feito por meio do script de envio já setado no código, mais especificamente no madata_package.

7.8 Processo de Empacotamento

Tendo em vista o escopo do projeto, e processo de utilização das funções e funcionalidades criadas, foi necessário realizar um processo de empacotamentos das funções e criação de uma biblioteca própria para que o cliente possa acessar e utilizar suas funcionalidade. Como contexto, o processo de empacotamento para projetos em Big Data tem como principais objetivos: a eficiência, reusabilidade e escalabilidade nos projetos.

Sendo assim, o processo foi dividido em algumas etapas:

- **Empacotamento de Código:**

- **Modularização:** Para facilitar a manutenção e o desenvolvimento, realizamos a modularização dos códigos em diferentes scripts e com orientação à objetos. Sendo assim, cada script possuía uma(s) função específica, facilitando a compreensão e a colaboração.
 - Especificamente neste caso, adotamos a parte de classes para que haja mais facilidade nessa manipulação das funções

- **Criação da Biblioteca:**

- **Encapsulamento das Funções:** As funções são encapsuladas dentro da biblioteca. Possuindo assim métodos reutilizáveis que podem ser chamados em diferentes partes do projeto, promovendo a reusabilidade do código.
 - Organização Lógica: isso facilita a organização lógica do código, agrupando funcionalidades similares.

- **Ferramentas da Biblioteca:**

- **Setuptools e Pip:** O `setuptools` é uma biblioteca Python que facilita a criação de pacotes Python. Ele é usado para empacotar a biblioteca, e em Python. O arquivo `setup.cfg` é utilizado para configurar os metadados do projeto e definir as instruções sobre como a biblioteca deve ser empacotada e instalada. O `pip` é uma ferramenta usada para instalar pacotes Python. Ele é usado para instalar as dependências necessárias para a biblioteca e para fazer o upload da biblioteca para o PyPI.
- **PySpark:** O PySpark é a versão Python do Apache Spark, uma estrutura de processamento de dados em paralelo de código aberto. Ele é usado para processar grandes conjuntos de dados e realizar análises complexas. No pacote MCDATA_PACKAGE, o PySpark é usado para transformar arquivos CSV e RDS para o formato Parquet.
- **Pandas:** O Pandas é uma biblioteca Python que é usada para manipulação e análise de dados. Ela é usada para manipular tabelas de dados no pacote.
- **Boto3:** O Boto3 é a biblioteca Amazon Web Services (AWS) Software Development Kit (SDK) para Python. Ela é usada para interagir com o serviço de armazenamento de objetos da AWS, o S3.
- **Twine:** O Twine é uma ferramenta usada para fazer o upload de pacotes Python para o PyPI.
- **PyPI:** O PyPI (Python Package Index) é um repositório online de software para Python. Ele é usado para distribuir a biblioteca para outros usuários.

- **Distribuição e Versionamento:**

- **Repositórios e Versionamento:** A biblioteca é distribuída por meio do repositório, PyPI. O versionamento, é gerenciado pelo `setup.config`, e

garante que diferentes versões da biblioteca possam ser controladas e instaladas conforme necessário.

- Para obter mais informações sobre procedimentos e configuração da biblioteca, acesse:

https://github.com/2023M8T4Inteli/grupo5/blob/dev/src/mcdata_package/README.md

<https://pypi.org/project/mcdata-package/>

8. Data Lake/Data Warehouse Alimentado pelo ETL

Neste documento explicaremos um pouco sobre o processo de inserção e formatação dos dados dentro do Datawarehouse. Cada uma das seções explicará um tópico específico deste processo contínuo desenvolvido na Sprint 3.

8.1 Fontes de dados

A razão para a escolha das fontes de dados utilizadas é a construção da possibilidade de se gerar uma visão abrangente e detalhada sobre o poder aquisitivo de diferentes regiões, sobre diferentes canais de venda ao redor do país e sobre os mercados e suas dinâmicas em cada lugar. Essas fontes devem permitir juntas a construção de análises úteis para a consultoria que trabalha com Go To Market e para a

qual seria valioso comparar informações relacionadas a estes assuntos. Segue descrição das fontes de dados utilizados:

Dados da ANVISA → possibilita a análise de poder aquisitivo por região, uma vez que exerce a vigilância sanitária sobre produtos e serviços no âmbito nacional. Duas tabelas foram utilizadas: uma de dados abertos sobre alimentos e uma sobre petições alimentícias. Estes dados possibilitam a construção de uma visão sobre hábitos de consumo e sobre outros aspectos relacionados a alimentos em diferentes regiões.

Dados do Bacen → O BACEN permite o acesso a informações sobre taxa selic, inflação e o valor da moeda em dólar. A coleta destes dados é essencial para que sejam traçadas visões gerais sobre a economia nas regiões de diferentes clientes.

Dados de CNPJ → O CNPJ armazena informações cadastrais das pessoas jurídicas e de outras entidades. Estes dados possibilitam um mapeamento granular dos diferentes canais de venda que habitam os mercados em regiões diversas.

Dados da POF → O POF (Pesquisa de Orçamentos Familiares) visa analisar as condições de vida da população brasileira a partir da análise de seus orçamentos.

Dados do IBGE → Instituto Brasileiro de Geografia e Estatística, analisa o território, a população e como a economia evolui através da produção e do consumo, revelando ainda aspectos de condições de vida das pessoas.

Ambas as fontes, POF e IBGE em geral, possibilitam o acesso a informações relacionadas ao poder aquisitivo de cada região na medida em que apresentam dados sobre a composição dos orçamentos domésticos e sobre as condições de vida da população.

Dados coletados com a API → Por fim, a API do cliente disponibiliza um CSV com todas as vendas realizadas por CNPJ, apresentando a data, o valor e a quantidade atrelados a cada um. Isso é essencial, uma vez que permite uma visualização objetiva das dinâmicas que ocorrem em diferentes mercados, tanto na questão econômica quanto na questão estritamente sobre consumo. É importante salientar que este dados são referentes a uma base de vendas fictícia fornecidas pela Integration.

8.1.1 Premissa de inclusão das Novas Fontes (Dados do Bacen e Dados da Anvisa)

A partir da revisão da última sprint e feedbacks passados pelo cliente, foi de comum acordo que havia necessidade de pensar diferente e incluir dados que poderiam trazer insights valiosos, e que ainda não foram incluídos e utilizados pela Integration. Neste sentido, trouxemos mais duas fontes de dados, no âmbito financeiro e outros dados referentes à Agência Nacional de Vigilância Sanitária.

8.1.1.1 Dados da Anvisa

A ANVISA (Agência Nacional de Vigilância Sanitária) disponibiliza dados abertos relacionados à petição de alimentos e dados referentes a alimentos e seus termos regulatórios. Esses dados podem ser valiosos para empresas no setor alimentício ao tomar decisões estratégicas, especialmente no contexto de go-to-market (inserido na presença da Integration e os insights tirados). Abaixo estão alguns pontos-chave que destacam a importância desses dados e seu impacto nas estratégias de go-to-market:

- 1. Segurança Alimentar e Conformidade:** Os dados de petição de alimentos da ANVISA podem fornecer informações sobre a conformidade de produtos alimentícios com regulamentações e padrões de segurança alimentar.

- Empresas podem utilizar esses dados para garantir que seus produtos atendam aos requisitos regulatórios, evitando problemas legais e protegendo a reputação da marca.
- 2. Avaliação da Competição:** Ao analisar as petições de alimentos submetidas por outras empresas do setor, é possível obter insights sobre as estratégias e inovações dos concorrentes. Isso permite que as empresas ajustem suas próprias estratégias de go-to-market para se destacarem no mercado e oferecerem produtos mais alinhados com as demandas dos consumidores.
- 3. Desenvolvimento de Produtos:** Os dados de petição de alimentos podem ser uma fonte valiosa para identificar tendências de mercado e demandas dos consumidores. As empresas podem utilizar essas informações para orientar o desenvolvimento de novos produtos que atendam às necessidades específicas do mercado.
- 4. Gerenciamento de Riscos:** Acompanhar as petições de alimentos pode ajudar as empresas a identificar potenciais riscos e problemas de segurança alimentar no mercado. Isso permite uma abordagem proativa na gestão de riscos, minimizando impactos negativos na reputação da empresa.
- 5. Tomada de Decisões Estratégicas:** Com base nos dados da ANVISA, as empresas podem tomar decisões mais informadas sobre sua presença no mercado. Isso inclui decisões relacionadas à expansão geográfica, ajustes no portfólio de produtos e aprimoramento de estratégias de marketing e distribuição.
- 6. Transparência e Confiança do Consumidor:** Utilizar dados de petição de alimentos e seguir as práticas recomendadas pela ANVISA pode aumentar a transparência das operações de uma empresa. Isso, por sua vez, pode fortalecer a

confiança dos consumidores, pois eles percebem que a empresa está comprometida com a segurança e conformidade alimentar.

8.1.1.2 Dados do Bacen

A inclusão da fonte de dados com informações do Banco Central do Brasil (BACEN), como a taxa Selic, a expectativa de inflação e o preço do dólar, pode proporcionar uma visão mais abrangente e estratégica para o cliente. Aqui estão algumas dessas oportunidades, no contexto do projeto:

1. Custo de Produção e Impacto Econômico:

- **Taxa Selic e Inflação:** A taxa Selic e a expectativa de inflação podem influenciar os custos de produção, financiamento e investimento. Monitorar esses indicadores ajuda as empresas a entenderem os impactos econômicos gerais em suas operações.

2. Cadeia de Suprimentos e Custos Logísticos:

- **Preço do Dólar:** O preço do dólar afeta diretamente os custos de importação de ingredientes e matérias-primas. Uma variação no câmbio pode impactar os custos logísticos e, consequentemente, os preços finais dos produtos.

3. Decisões de Precificação e Estratégias de Mercado:

- **Inflação e Câmbio:** A expectativa de inflação e o preço do dólar podem influenciar as estratégias de precificação. As empresas podem ajustar os preços de acordo com as condições econômicas para manter a competitividade e a rentabilidade.

4. Investimentos em Pesquisa e Desenvolvimento:

- **Taxa Selic:** Mudanças na taxa Selic podem afetar os custos de capital e financiamento para projetos de pesquisa e desenvolvimento. Acompanhar a taxa Selic é crucial para avaliar o ambiente de investimento.

5. Planejamento Estratégico e Expansão de Mercado:

- **Câmbio e Inflação:** Flutuações cambiais e expectativas de inflação podem influenciar as estratégias de expansão para novos mercados. As empresas podem ajustar suas abordagens com base nas condições macroeconômicas.

6. Gestão de Riscos e Estabilidade Financeira:

- **Taxa Selic e Câmbio:** A taxa Selic pode impactar o custo do capital, enquanto as variações cambiais podem representar riscos financeiros. Uma gestão eficaz desses riscos é crucial para a estabilidade financeira.

7. Compreensão do Consumidor e Estratégias de Marketing:

- **Inflação:** A expectativa de inflação pode influenciar o poder de compra dos consumidores. Compreender essas tendências é essencial para ajustar estratégias de marketing e comunicação.

8.2 Etapas do ETL

O ETL (Extração, Transformação e Carga) envolve a coleta de dados de diversas fontes, a aplicação de transformações para garantir qualidade e consistência, e o carregamento dos dados em um destino, neste caso, um data warehouse. O ETL é necessário para a conversão de grandes volumes de dados brutos em informações valiosas, facilitando a inteligência de negócios e a tomada de decisões informadas.

Explicação mais detalhada:

Extração

- A Extração é a primeira fase do processo ETL, onde os dados são coletados de várias fontes que podem variar desde bancos de dados até serviços na nuvem. A extração deve ser projetada de forma a minimizar o impacto no desempenho dos sistemas de origem e deve ser eficiente em termos de recursos para lidar com o volume de dados extraídos dentro do tempo disponível.

Transformação

- A Transformação é a segunda fase do ETL. Durante esta etapa, os dados extraídos são limpos, validados e transformados para se adequarem ao esquema do data warehouse de destino. As transformações podem incluir a conversão de tipos de dados ou a aplicação de funções de negócios para calcular métricas de negócios.

Carga

- A Carga é a última fase do ETL. Nesta etapa, os dados transformados são carregados no data warehouse de destino. O processo de carga também deve garantir que os dados sejam carregados de forma eficiente e que o data warehouse permaneça disponível e estável durante o processo de carga.

Na nossa solução, a etapa de extração foi realizada a partir do DataLake, especificamente do AWS S3, utilizando os buckets que compõem esse DataLake, como o bucket "grupo-5-mec-solutions-bucket-dados-financeiros-conta-nova". A extração dos dados foi feita diretamente pelo AWS Redshift, permitindo que todas as tabelas localizadas nos buckets possam ser manipuladas e transformadas.

As transformações realizadas nos dados extraídos incluíram a modificação das informações das colunas para torná-las mais visuais e compreensíveis, como a mudança do número do estado para o nome do estado (11 → São Paulo). Outro exemplo foi normalização das colunas, para que todas tabelas com colunas semelhantes em csvs distintos tivessem o mesmo nome.

Mais alguns tratamentos feitos:

```
1 ✓ CREATE TABLE public.anvisa (
2     num_expediente_peticao bigint ENCODE az64,
3     num_processo_peticao bigint ENCODE az64,
4     s_n_peticao_primaria character varying(256) ENCODE lzo,
5     cod_assunto_peticao character varying(256) ENCODE lzo,
6     desc_assunto_peticao character varying(300) ENCODE lzo,
7     data_situacao_atual_peticao timestamp without time zone ENCODE az64,
8     desc_situacao_atual_peticao character varying(256) ENCODE lzo,
9     data_primeira_finalizacao timestamp without time zone ENCODE az64,
10    data_finalizacao_atual timestamp without time zone ENCODE az64,
11    desc_tipo_documento character varying(256) ENCODE lzo,
12    desc_area_interesse character varying(256) ENCODE lzo,
13    desc_fila_analise character varying(256) ENCODE lzo,
14    desc_sub_fila_lista_analise character varying(256) ENCODE lzo,
15    desc_grupo_etapa_ciclo_analise character varying(256) ENCODE lzo,
16    data_ini_ocorrencia_grp_etapa timestamp without time zone ENCODE az64,
17    data_fim_ocorrencia_grp_etapa timestamp without time zone ENCODE az64,
18    ordem_ocorre_grupo_etapa_asc integer ENCODE az64,
19    ordem_ocorre_grupo_etapa_desc integer ENCODE az64
20 ) DISTSTYLE AUTO;
```

Load data

Frequently used parameters

Ignore header rows (IGNOREHEADER)

Treats the specified number_rows as a file header and doesn't load them. Use this option to skip file headers in all files in a parallel load.

1



Time format (TIMEFORMAT)

Specifies the time format. If no format string is specified, the default format is YYY-MM-DD HH:MI:SS for TIMESTAMP columns or YYYY-MM-DD HH-MI-SSOF for TIMESTAMPTZ columns, where OF is the offset from Coordinated Universal Time (UTC).

auto

epochsecs

epochmillisecs

format

Time format string

Date format (DATEFORMAT)

Specifies the date format. If no format string is specified, the default format is 'YYY-MM-DD'. If data import doesn't recognize the format of your date values, or if your date values use different formats, select the 'auto' option.

auto

format

Date format string

Accept characters that aren't valid (ACCEPTINVCHARS)

When this option is specified, COPY replaces each invalid UTF-8 character with a string of equal length consisting of the replacement characters.

Replacement character

Load data

Load blank fields as NULL (BLANKSASNNULL)

Loads blank fields, which consist of only white space characters, as NULL. This option applies only to CHAR and VARCHAR columns.

Load empty char fields as null (EMPTYASNNULL)

Indicates that Amazon Redshift should load empty CHAR and VARCHAR fields as NULL.

Encoding (ENCODING)

Specifies the encoding type of the load data. The COPY command converts the data from the specified encoding into UTF-8 during loading.

UTF8 UTF16 UTF16LE UTF16BE

Explicit identity columns (EXPLICIT_IDS)

Use this option with tables that have IDENTITY columns if you want to override the autogenerated values with explicit values from the source data files for the tables.

Fill missing columns (FILLRECORD)

Allows data files to be loaded when contiguous columns are missing at the end of some of the records.

Ignore blank lines (IGNOREBLANKLINES)

Ignores blank lines that only contain a line feed in a data file and does not try to load them.

Remove quotation marks (REMOVEQUOTES)

Removes surrounding quotation marks from strings in the incoming data.

Round up numeric values (ROUNDDEC)

Rounds up numeric values when the scale of the input value is greater than the scale of the column.

Back

Cancel

Next

Por fim, na fase de carga, novamente utilizamos o AWS Redshift para incorporar os dados processados pelo ETL no data warehouse, garantindo assim uma organização ordenada. Este ciclo completo do ETL é fundamental para manter a integridade e a utilidade dos dados, proporcionando uma base robusta para análises avançadas e decisões estratégicas bem fundamentadas.

8.3 Serviços de ETL e armazenamento

Escolhemos o Amazon Redshift como nossa solução principal para ETL e armazenamento de dados, principalmente devido às suas características orientadas para processamento analítico online (OLAP) e a sua conexão fácil com o AWS S3 (datalake), permitindo uma extração facilitada dos dados. Outro ponto que consideramos foi a escalabilidade dos serviços e sua utilização.

Alguns pontos que também foram importantes para a escolha desse serviço foram:

Arquitetura colunar: Ao contrário de bancos de dados relacionais tradicionais que armazenam dados em linhas, os bancos de dados colunares organizam os dados em colunas e, assim, otimiza as consultas OLAP, pois permite a leitura seletiva das colunas relevantes, minimizando o tempo de busca e aumentando o desempenho geral.

Computação distribuída: Em vez de depender de um único sistema centralizado para realizar todas as tarefas, a carga de trabalho é distribuída entre vários dispositivos para melhorar a eficiência, escalabilidade e confiabilidade do sistema como um todo. → “O Redshift é 10 vezes mais rápido do que Hadoop. Em alguns testes de consulta, o banco de dados Redshift supera facilmente o Hadoop ao retornar resultados.” (**5 benefícios em usar o Redshift para Data Warehouse**)

Monitoramento de clusters: O Redshift oferece várias ferramentas de monitoramento, permitindo rastrear a integridade e o desempenho dos clusters e bancos de dados (será abordado mais a frente no documento) → “O [Redshift](#) apresenta algumas ferramentas diferentes de criptografia e segurança que tornam a proteção de depósitos ainda mais fácil.” (**5 benefícios em usar o Redshift para Data Warehouse**)

Essas características fazem do Amazon Redshift a nossa escolha de datawarehouse, proporcionando escalabilidade, desempenho e facilidade de uso.

8.4 Descrição da estrutura do Data Warehouse

O Data Warehouse é uma ferramenta fundamental para organizar e disponibilizar informações para a tomada de decisões estratégicas. Entre as diversas plataformas disponíveis, destaca-se o Amazon Redshift, um sistema de gerenciamento de banco de dados especialmente projetado para armazenar e analisar grandes conjuntos de dados. Ele foi o serviço escolhido para construção do datawarehouse.

Neste contexto, exploraremos a estrutura do Data Warehouse no ambiente OLAP (Online Analytical Processing), onde a ênfase recai sobre a capacidade de analisar dados multidimensionais de maneira eficiente.

Primeiramente, é preciso entender os dois diferentes tipos de processamento que foram considerados na possível agregação na arquitetura de ingestão de dados. O OLAP (Online Analytical Processing) e OLTP (Online Transaction Processing), desempenham papéis distintos e vitais no ecossistema da gestão de informações. Estas duas abordagens, embora compartilhem a mesma base de dados, são orientadas para propósitos específicos, refletindo suas características e funcionalidades únicas.

8.4.1 OLAP

O OLAP, centrado na análise, é utilizado com a intenção de extrair conhecimento significativo a partir de grandes conjuntos de dados. Sua estrutura multidimensional permite relações entre diferentes variáveis de forma intuitiva, utilizando cubos de dados organizados em dimensões. Esse sistema foi escolhido para ser utilizado no projeto por

suas características de inferência mais analítica e fácil processamento de dados complexos e grandes.

É importante citar que o ambiente em sistemas OLAP é particularmente eficaz para consultas pesadas e relatórios estratégicos, oferecendo uma visão abrangente do desempenho organizacional e padrões no cubo de dados.

8.4.2 OLTP

Em contraste, o OLTP é otimizado para o processamento eficiente de transações em tempo real. É muito utilizado em ambientes transacionais, como sistemas de gerenciamento de banco de dados operacionais. No contexto deste sistema o foco recai sobre a rapidez e a precisão na execução de operações individuais, como inserção, atualização e exclusão de registros. A normalização é frequentemente aplicada para garantir a integridade dos dados, evitando redundâncias e mantendo a consistência em cenários dinâmicos.

8.4.3 Mais diferenças

Para entender melhor sua aplicabilidade, é importante diferenciar os conceitos e utilização desses sistemas em arquiteturas de Big Data. Uma das principais distinções reside na maneira como esses sistemas tratam a concorrência de acesso aos dados. Enquanto o OLAP prioriza a leitura simultânea de dados por vários usuários, o OLTP é projetado para lidar com a concorrência de gravação, garantindo a consistência dos dados em ambientes onde múltiplas transações ocorrem simultaneamente. As necessidades de armazenamento e recuperação de dados também diferem substancialmente entre OLAP e OLTP. O OLAP frequentemente emprega estratégias de armazenamento de dados em formato desnormalizado, visando otimizar o desempenho das consultas analíticas. Por

outro lado, o OLTP favorece abordagens normalizadas para minimizar o espaço de armazenamento e assegurar a integridade dos dados nas transações diárias.

Com intuito de uma visualização mais intuitivas, representamos as características de cada um dos sistemas, destacando o OLAP que foi o escolhido para a arquitetura de ingestão:

Sistemas de processamento de dados



O Amazon Redshift, oferece também uma abordagem eficiente para a organização de dados em um ambiente analítico. Em sua construção ele faz a formatação dos dados e em sua entrada posiciona-os de maneira relacional, utilizando um sistema de arquivos distribuídos para otimizar o acesso e a manipulação de informações.

8.4.4 Organização Relacional:

A estrutura relacional do Amazon Redshift é fundamentada nos princípios da modelagem dimensional. Os dados são organizados em tabelas, seguindo esquemas como o star schema ou snowflake schema. O star schema, por exemplo, envolve uma tabela fato central, que contém as métricas a serem analisadas, e tabelas de dimensões que

fornecem contexto aos dados. Isso simplifica as consultas, permitindo que os usuários explorem relações multidimensionais de maneira eficiente.

As tabelas são distribuídas em "slices" dentro dos nós de processamento do cluster do Redshift. Cada nó é uma unidade de computação independente, e a distribuição das tabelas entre esses nós é projetada para otimizar o desempenho das consultas. O Redshift permite a escolha do método de distribuição mais adequado para cada tabela, oferecendo flexibilidade para adaptar a arquitetura à natureza dos dados e dos padrões de consulta.

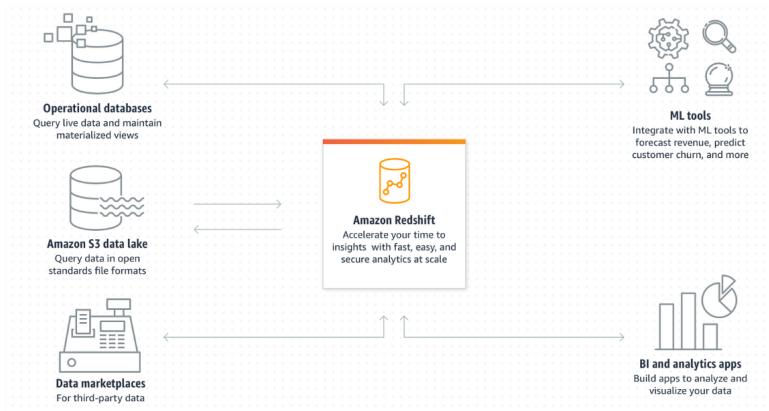
8.4.5 Sistema de Arquivos Distribuídos:

Outro característica muito legal do Redshift é seu processamento distribuído. O sistema armazena dados em um sistema de arquivos distribuídos, onde os dados de cada tabela são particionados em blocos e distribuídos entre os nós do cluster. Cada nó é capaz de processar consultas localmente nos dados armazenados em seu próprio espaço de armazenamento, minimizando a necessidade de transferência de dados entre os nós durante a execução de consultas.

A estrutura de coluna do Redshift também desempenha um papel importante na otimização do armazenamento e recuperação de dados. Os dados são armazenados de forma colunar em vez de linha, o que permite a compressão eficiente e a leitura seletiva de colunas durante as consultas, resultando em tempos de resposta mais rápidos (isso pode ser evidenciado nas métricas dos alarmes em cada serviço salientadas no tópico "Proposta de monitoramento e gerenciamento do ETL" desta docume

Organização do Datawarehouse

Posicionamento dos dados de forma relacional e organizando-os num sistema de arquivos distribuídos:

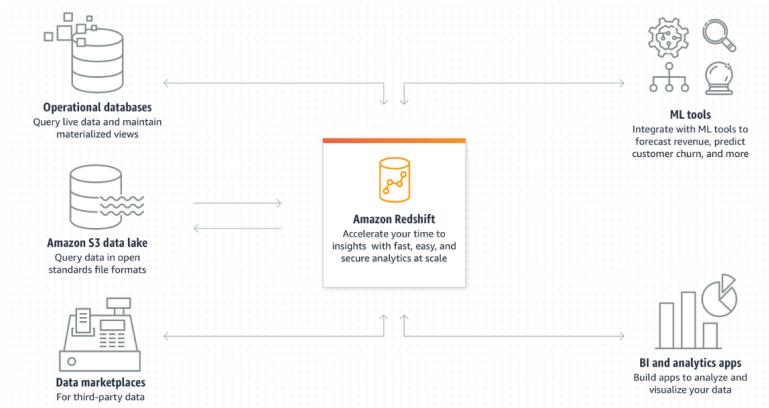


O Redshift é um serviço OLAP e tem processamento distribuído

Informações do Ambiente

Organização do Datawarehouse

Posicionamento dos dados de forma relacional e organizando-os num sistema de arquivos distribuídos:



O Redshift é um serviço OLAP e tem processamento distribuído

- **Namespace:** aula-afonso
- **ID do Namespace:** 3af09ae1-e265-4a72-b980-796bfef5b657

- **ARN do Namespace:**

arn:aws:redshift-serverless:us-east-1:187967616478:namespace/3af09ae1-e
265-4a72-b980-796bfef5b657

- **Data de Criação:** 21 de Novembro, 2023, 16:59 (UTC-03:00)
- **Usuário Admin:** admin
- **Nome do Banco de Dados:** dev
- **Armazenamento Utilizado:** 27,2 GB
- **Contagem Total de Tabelas:** 12

Snapshot

O snapshot como uma cópia de um cluster do Redshift, os snapshots são ferramentas valiosas para a recuperação de dados, backup e manutenção da integridade do ambiente de armazenamento. Esses instantâneos fornecem uma maneira eficaz de preservar o estado consistente dos dados, permitindo que as organizações restaurem clusters a um ponto anterior no tempo em caso de falhas, erros ou necessidades de análise histórica.

Segue detalhes e comprovação da criação do Snapshot:

Snapshots (1) [Informações](#)

Selecione pelo menos um snapshot para realizar uma ação.

<input type="checkbox"/>	Snapshot	Status	Namespace	Hora de criação	Tamanho dos dados	Tags	Tempo até a exclusão	ARN do snapshot
<input type="checkbox"/>	snap1	Disponível Available	aula-afonso	November 23, 2023, 17:52 (UTC-03:00)	22,6 GB	0 tag	Retido indefinidamente	arn:aws:redshift-serverless:us-east-

Detalhes do snapshot

Identificador de snapshot snap1	Status Available
ARN do snapshot arn:aws:redshift-serverless:us-east-1:187967616478:snapshot/727dd970-c87c-4c96-bdf4-96db673edfc5	

Detalhes do backup

Tamanho total do backup 22,6 GB	Período de retenção Indefinidamente
Criptografado Verdadeiro	Tempo até a exclusão Retido indefinidamente
Data de criação November 23, 2023, 17:52 (UTC-03:00)	Progresso  Backup de 22,6 GB bem-sucedido

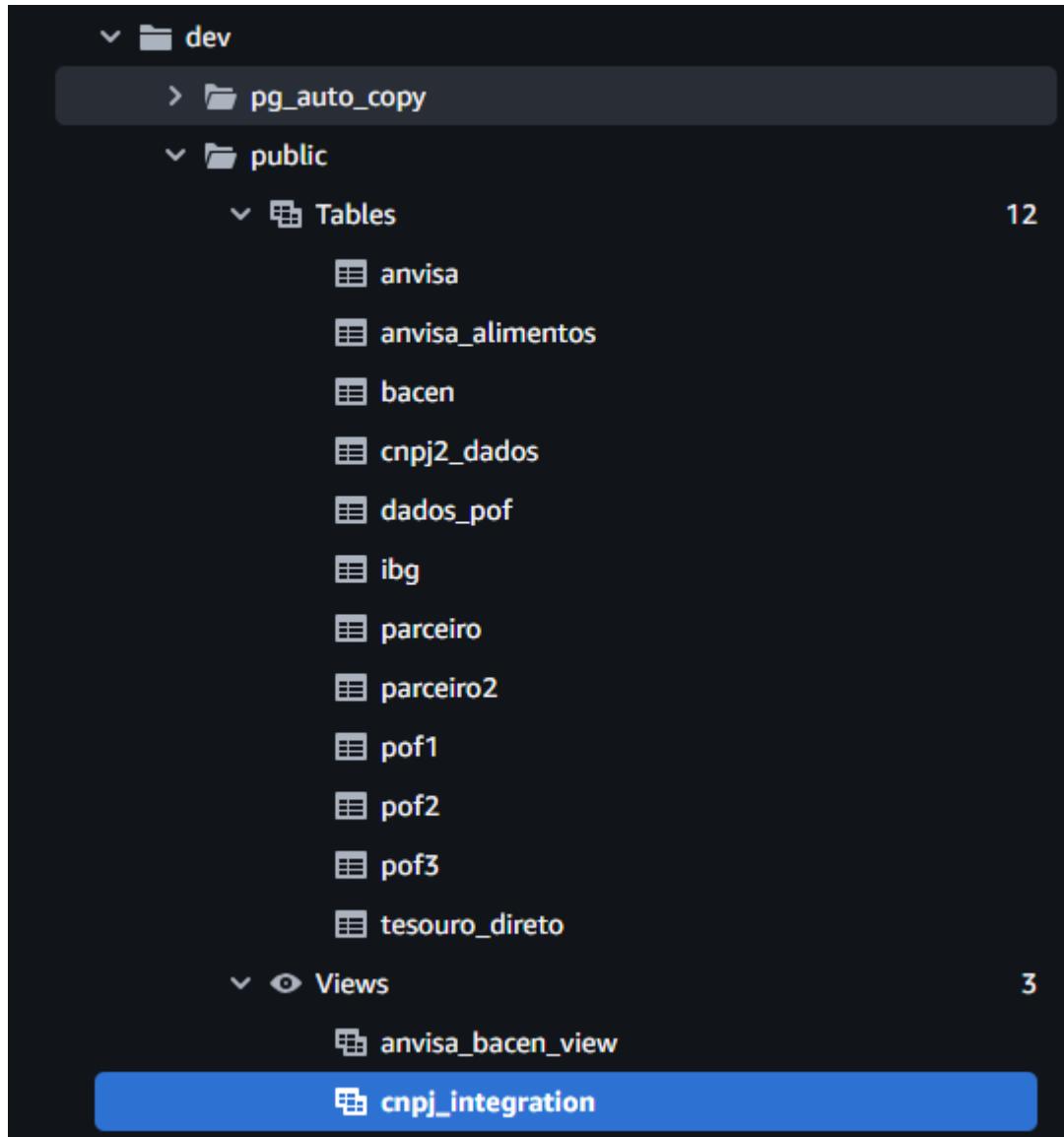
Namespace details

Namespace aula-afonso	Status Available
Namespace ID 3af09ae1-e265-4a72-b980-796bfef5b657	Data de criação November 21, 2023, 16:59 (UTC-03:00)
Namespace ARN arn:aws:redshift-serverless:us-east-1:187967616478:namespace/3af09ae1-e265-4a72-b980-796bfef5b657	Storage used 27,2 GB

8.5 Views + Tabelas do Redshift

As views para um contexto de definição e estruturação do Datawarehouse desempenham um papel muito importante, agindo como representações virtuais de tabelas derivadas de uma ou mais fontes de dados. Estas views não armazenam dados fisicamente, mas facilitam a consulta e manipulação dos dados de forma simplificada e eficiente.

Segue Organização das Tabelas e Views no Redshift:



8.5.1 O que é?

As Views em SQL, podem ser definidas como consultas armazenadas que funcionam como tabelas virtuais. Elas são construídas com base em consultas SQL, permitindo aos usuários acessar e manipular dados de maneira simplificada (*Material Didático - IMD*, n.d.).

8.5.2 Para que serve?

As views têm uma variedade de propósitos em um banco de dados SQL. Elas são utilizadas para simplificar a complexidade do esquema, oferecendo uma camada de abstração sobre as tabelas subjacentes. Além disso, as views são valiosas para a segurança de dados, permitindo a imposição de políticas de restrição de acesso. Elas facilitam a personalização da apresentação dos dados, agregação e resumo, contribuindo para a eficiência na análise e tomada de decisões. As views também desempenham um papel importante na padronização da visualização dos dados, promovendo uma interpretação consistente dentro da organização.

8.5.3 Benefícios

Um dos principais benefícios das views é a segurança de dados. Elas possibilitam a imposição de políticas de segurança, restringindo o acesso dos usuários apenas às informações pertinentes às suas funções, garantindo assim a integridade e confidencialidade dos dados. Além disso, as views permitem personalização da visualização, permitindo que os usuários organizem e apresentem os dados de acordo com suas necessidades específicas. Isso é particularmente útil em ambientes nos quais diferentes partes da organização precisam interpretar os mesmos dados de maneiras distintas.

A capacidade de agregação e resumo é outra vantagem das views, possibilitando a criação de visualizações que resumem informações, facilitando a identificação de tendências e padrões. Esse recurso é essencial para uma análise eficaz, especialmente em Data Warehouses que lidam com grandes volumes de dados. Além de contribuir para a eficiência na análise, as views também podem ser otimizadas para melhorar o desempenho das consultas, o que é crucial em ambientes de Data Warehouse nos quais

a eficiência no acesso aos dados é fundamental. A padronização da apresentação é promovida pela criação de views padronizadas, o que garante uma representação consistente dos dados e facilita a comunicação dentro da organização. Isso cria uma compreensão uniforme da estrutura do banco de dados e dos relacionamentos entre as tabelas.

No âmbito prático, a utilização de views não apenas simplifica a interpretação de dados complexos, mas também suporta a tomada de decisões informadas. Ao fornecer uma visão clara e personalizada dos dados, as views capacitam os usuários a tomar decisões estratégicas com base nas informações disponíveis no Data Warehouse.

Em resumo, as views desempenham um papel fundamental na maximização do valor dos dados armazenados em um Data Warehouse, facilitando a compreensão, análise e utilização eficaz das informações pelos usuários.

8.5.4 Como foi aplicado?

Dado o contexto do projeto, projetamos algumas views de exemplo para começar a visualizar a correlação e definir algumas premissas iniciais para tomada de decisão na parte da identificação de padrões.

A primeira view criada foi realizada com o intuito de visualizar relações entre dados de petição de alimentos da ANVISA e dados do Banco Central (BACEN). Neste caso foi criado para que simplificasse a análise em relação a estes dados. Foi construída também para mostrar a relação entre petições de alimentos e indicadores econômicos, como a taxa Selic e o preço do dólar. Isso permite uma visão abrangente no setor alimentício, com insights adicionais como fatores econômicos que impactam as petições e permitem que a empresa adaptem suas estratégias.

8.5.4.1 Dados Anvisa vs Dados do Bacen

desc_sub_fila_lista_an...	desc_grupo_etapa_cicl...	data_ini_ocorrencia_gr...	data_fim_ocorrencia_g...	ordem_ocorre_grupo_...	ordem_ocorre_grupo_...	data	usd
PETI??ES SIMPLIFICADAS	An?lise em Andamento	2007-01-26 00:00:00	2007-02-05 00:00:00	7	5	2007-02-05	2.0956
PETI??ES SIMPLIFICADAS	Sobrestado Externo	2007-02-05 00:00:00	2007-02-07 00:00:00	8	4	2007-02-07	2.0844
PETI??ES SIMPLIFICADAS	An?lise em Andamento	2007-02-07 00:00:00	2007-02-07 00:00:00	9	3	2007-02-07	2.0844
PETI??ES SIMPLIFICADAS	An?lise de Cumprimento ...	2007-02-07 00:00:00	2007-02-07 00:00:00	10	2	2007-02-07	2.0844
PETI??ES SIMPLIFICADAS	Finaliza??o	2007-02-07 00:00:00	2007-02-12 00:00:00	11	1	2007-02-12	2.1132
PETI??ES SIMPLIFICADAS	Fila de An?lise	2006-06-02 00:00:00	2006-06-08 00:00:00	1	7	2006-06-08	2.2693
PETI??ES SIMPLIFICADAS	An?lise em Andamento	2006-06-08 00:00:00	2006-12-20 00:00:00	2	6	2006-12-20	2.1552
PETI??ES SIMPLIFICADAS	Exig?ncia	2006-12-20 00:00:00	2007-01-19 00:00:00	3	5	2007-01-19	2.1299
PETI??ES SIMPLIFICADAS	An?lise de Cumprimento ...	2007-01-19 00:00:00	2007-01-19 00:00:00	4	4	2007-01-19	2.1299
PETI??ES SIMPLIFICADAS	An?lise em Andamento	2007-01-19 00:00:00	2007-01-31 00:00:00	5	3	2007-01-31	2.1239
PETI??ES SIMPLIFICADAS	An?lise de Cumprimento ...	2007-01-31 00:00:00	2007-01-31 00:00:00	6	2	2007-01-31	2.1239
PETI??ES SIMPLIFICADAS	Finaliza??o	2007-01-31 00:00:00	2007-02-05 00:00:00	7	1	2007-02-05	2.0956
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
TRANSFER?NCIA DE TI...	Finaliza??o	2006-06-07 00:00:00	2006-06-12 00:00:00	2	1	2006-06-12	2.2699
PETI??ES SIMPLIFICADAS	Sobrestado Externo	2005-11-21 00:00:00	2005-11-22 00:00:00	2	3	2005-11-22	2.2503
PETI??ES SIMPLIFICADAS	An?lise em Andamento	2005-11-22 00:00:00	2005-11-28 00:00:00	3	2	2005-11-28	2.2086
PETI??ES SIMPLIFICADAS	Finaliza??o	2005-11-28 00:00:00	2005-12-05 00:00:00	4	1	2005-12-05	2.196
PETI??ES SIMPLIFICADAS	Sobrestado Externo	2005-11-21 00:00:00	2005-11-22 00:00:00	2	3	2005-11-22	2.2503
PFTI??FS SIMPLIFICADAS	An?lise em Andamento	2005-11-22 00:00:00	2005-11-28 00:00:00	3	2	2005-11-28	2.2086

Query breakdown:

```

● ● ●

CREATE VIEW public.anvisa_bacen_view AS
SELECT
    a.*,
    b.*
FROM
    public.anvisa a
JOIN public.bacen b ON a.data_fim_ocorrencia_grupo_etapa = b.data;

```

```

CREATE
OR REPLACE VIEW "public"."anvisa_bacen_view" AS
SELECT
    a.num_expediente_peticao,
    a.num_processo_peticao,
    a.s_n_peticao_primaria,
    a.cod_assunto_peticao,
    a.desc_assunto_peticao,
    a.data_situacao_atual_peticao,
    a.desc_situacao_atual_peticao,
    a.data_primeira_finalizacao,
    a.data_finalizacao_atual,
    a.desc_tipo_documento,
    a.desc_area_interesse,
    a.desc_fila_analise,
    a.desc_sub_fila_lista_analise,
    a.desc_grupo_etapa_ciclo_analise,
    a.data_ini_ocorrencia_grp_etapa,
    a.data_fim_ocorrencia_grp_etapa,
    a.ordem_ocorre_grupo_etapa_asc,
    a.ordem_ocorre_grupo_etapa_desc,
    b."data",
    b.usd,
    b.taxa_selic,
    b.taxa_exp_inflacao
FROM
    anvisa a
    JOIN bacen b ON a.data_fim_ocorrencia_grp_etapa = b."data":: timestamp without time zone;

```

Inferências e correlação identificadas por meio da criação da view acima:



*dados de 2005 apenas como exemplificação

Da mesma forma, relacionamos os dados de CNPJ com informações de vendas de uma empresa do setor alimentício provisionados pela API do cliente. A view podem ser

utilizada para criar uma visão consolidada que mostra a performance de vendas. Isso facilita a análise de padrões de vendas, identificação de clientes importantes e otimização de estratégias de marketing e distribuição.

8.5.4.2 Dados da API do Parceiro vs Dados de CNPJ

cnpj	idcategory	produto	data	preco	quantidade	cnpj_cnpj2_dados	nome_fantasia
34639031000109	5	Ketchup	2023-10-25	7	91	34639031000109	OVERCOME
4450754000102	20	Tênis	2023-01-27	46	5	4450754000102	TABERNA DO POR
35001876000137	3	Vinho	2023-11-20	256	37	35001876000137	CARIOCAS BEER
20266780000123	24	Livro	2023-10-14	633	42	20266780000123	ICEMELLOW
29991192000127	17	Queijo	2023-11-19	913	34	29991192000127	SOHFRUTA
37638065000177	2	Cerveja	2023-04-19	846	13	37638065000177	SCHEILA NASCIME
34639031000109	5	Ketchup	2023-10-25	7	91	34639031000109	OVERCOME
36316296000100	6	Oleo	2023-07-05	29	61	36316296000100	BIG DOG LANCHES
31905602000111	20	Tênis	2023-04-21	657	67	31905602000111	
43070699000179	1	Chocolate	2023-06-15	387	48	43070699000179	LA PASTA
41892024000180	1	Chocolate	2023-01-25	777	95	41892024000180	DON MARTONE
42133073000100	4	Arroz	2023-08-27	687	8	42133073000100	FAST LOUNGE BEB
43070699000179	1	Chocolate	2023-06-15	387	48	43070699000179	LA PASTA
39907998000148	12	Pão de Forma	2023-01-26	877	74	39907998000148	S.S. SALGADOS
41116949000139	10	Batata Palha	2023-06-25	659	31	41116949000139	MIXY - ASSESSORI
35304789000159	10	Batata Palha	2023-03-20	249	51	35304789000159	ZIAS BURGUER
37638065000177	2	Cerveja	2023-04-19	846	13	37638065000177	SCHEILA NASCIME
37125897000190	20	Tênis	2023-07-04	951	81	37125897000190	SUBMARINO BUR
26325602000102	27	Notebook	2023-01-04	400	84	26325602000102	MARE SUSHI OFIC
21799743000143	30	Carne	2023-05-23	79	97	21799743000143	DAY FIGUEIREDO I
4450754000102	20	Tênis	2023-01-27	46	5	4450754000102	TABERNA DO POR
40269366000185	6	Oleo	2023-11-22	966	77	40269366000185	LIDELICIAS MACAE
34028419000173	2	Cerveja	2023-04-13	435	87	34028419000173	BLACKBURGER

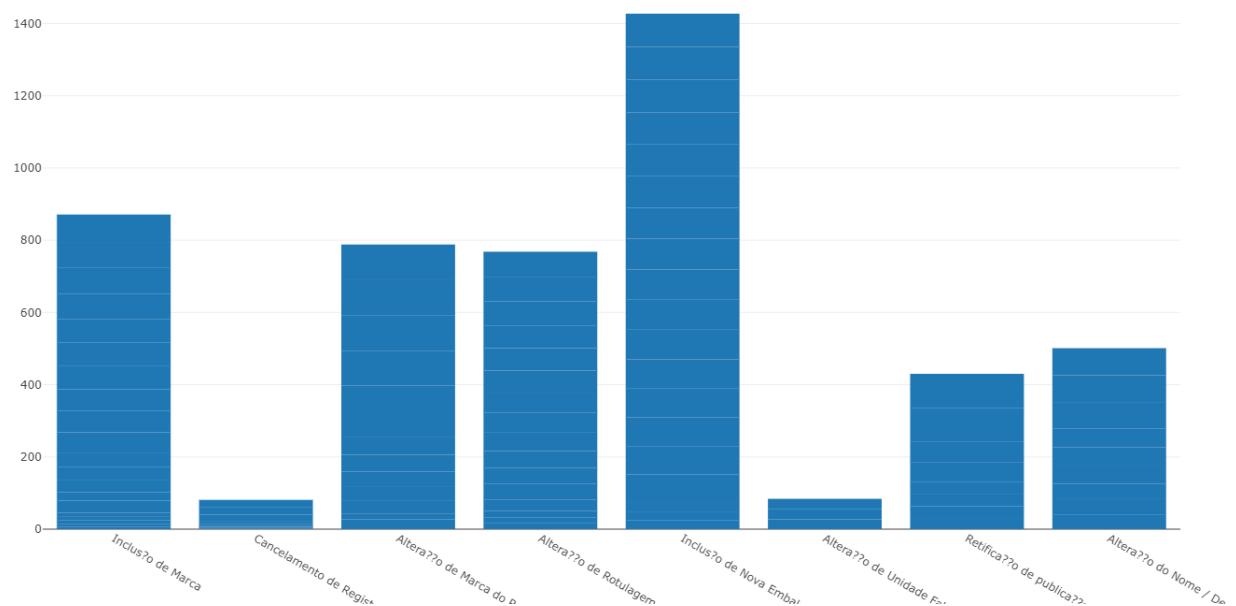
Query breakdown:

```

CREATE
OR REPLACE VIEW "public"."cnpj_integration" AS
SELECT
    y.id,
    y.cnpj,
    y.idcategory,
    y.producto,
    y."data",
    y.preco,
    y.quantidade,
    z.cnpj AS cnpj_cnpj2_dados,
    z.nome_fantasia,
    z.sigla_uf
FROM
    parceiro2 y
JOIN cnpj2_dados z ON y.cnpj = z.cnpj;

```

Inferências e correlação identificadas por meio da criação da view acima:



Por fim, é possível inferir que a criação de views oferece uma abordagem flexível e poderosa para visualizar e analisar relações entre dados, proporcionando uma visão personalizada e eficiente para suportar a tomada de decisões estratégicas.

8.6 Segurança / Privacidade / Conformidade

Para assegurar a integridade, segurança e conformidade regulatória dos dados durante o processo de ELT (Extract, Load, Transform), foram adotadas medidas específicas:

8.6.1 Segurança dos Dados:

- **Criptografia de Dados em Trânsito:** Utilização de protocolos de criptografia SSL/TLS durante a transferência de dados do Amazon S3 para o Amazon Redshift, garantindo que os dados sejam transmitidos de forma segura.
- **Controle de Acesso:** Implementação de políticas de controle de acesso granulares, baseadas em IAM (Identity and Access Management), para garantir que apenas usuários autorizados possam acessar os dados e realizar operações no Redshift.

Workgroup	Backup de dados	Segurança e criptografia	Unidades de compartilhamento de dados	Zero-ETL integrations	Resource policy	Tags
Permissões <small>Informações</small>						
Funções do IAM	Status	Nome do recurso da Amazon (ARN)		Tipo de função		
iam_redshift	in-sync	arn:aws:iam::187967616478:role/iam_redshift	-	-	-	
rodrigo-S3	in-sync	arn:aws:iam::187967616478:role/rodrigo-S3	-	-	-	

8.6.2 Privacidade dos Dados:

- **Dados Públicos ou Fictícios:** Considerando que os dados utilizados são públicos ou fictícios, não foi aplicada a prática de anonimização ou mascaramento de dados, uma vez que não contêm informações identificáveis ou sensíveis.

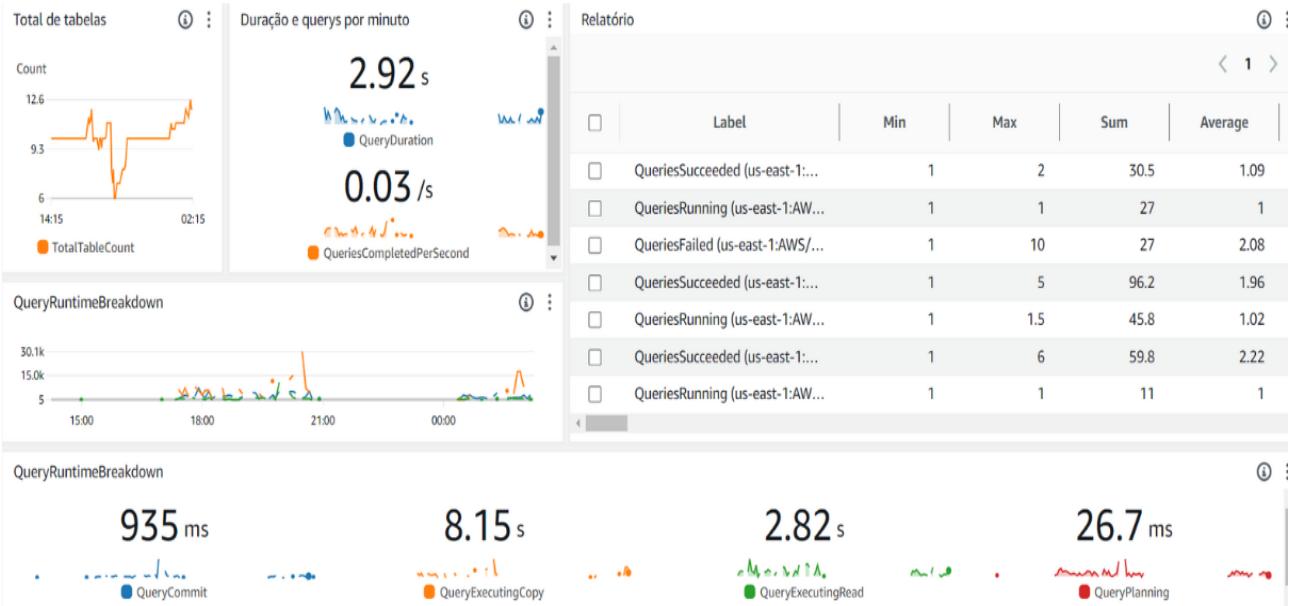
8.6.3 Backup de Segurança:

- **Snapshot:** Implementamos um procedimento de criação regular de snapshots dos dados no Amazon Redshift. Isso nos permite manter cópias de segurança dos dados, proporcionando uma camada adicional de segurança em caso de problemas ou necessidade de recuperação.

Snapshots (1) Informações						
Selecionar pelo menos um snapshot para realizar uma ação.						
<input type="checkbox"/>	Snapshot	Status	Namespace	Hora de criação	Tamanho dos dados	Tags
<input type="checkbox"/>	snap1	Disponível Available	aula-afonso	November 23, 2023, 17:52 (UTC-03:00)	22,6 GB	0 tag

Estas medidas foram implementadas considerando as melhores práticas de segurança e conformidade, adequadas ao contexto de dados públicos ou fictícios, assegurando a proteção dos dados e o atendimento aos requisitos regulatórios aplicáveis.

8.7 Proposta de monitoramento e gerenciamento do ETL



8.7.1 Gráficos

8.7.1.1 Total de tabelas:

Duração das querys e querys por minuto:

1. Duração das Queries:

Mede o tempo que uma consulta leva para ser executada. Ajuda a identificar e otimizar consultas lentas, melhorando a eficiência do banco de dados.

1. Queries por Minuto:

Indica o número de consultas executadas por minuto. Revela a carga de trabalho do banco de dados ao longo do tempo, facilitando ajustes de recursos conforme necessário.

Essas métricas são cruciais para manter o desempenho e a eficiência do banco de dados, mesmo sem o uso específico do Amazon CloudWatch. Utilize outras ferramentas de monitoramento ou as métricas internas do seu sistema de gerenciamento de banco de dados.

8.7.1.2 Relatório de querys:

O CloudWatch geralmente oferece métricas relacionadas ao desempenho, como CPU, I/O de disco, número de conexões e outras métricas específicas do banco de dados. Essas métricas podem ser úteis para avaliar o desempenho geral do banco de dados e identificar possíveis problemas.

8.7.1.2 QueryRuntimeBreakdown:

Métricas relacionadas ao desempenho de consultas podem incluir informações sobre o tempo de execução de consultas, detalhes sobre a distribuição do tempo gasto em diferentes componentes da consulta (como processamento no lado do servidor, transferência de dados, etc.) e outros fatores que podem impactar o desempenho geral do sistema.

8.7.2 Alarmes

No CloudWatch é possível configurar alarmes para casos diversos. Cada alarme contém os parâmetros nome, estado, última atualização de estado, condições e ações. O parâmetro “estado” define se o alarme está desativado, se está “ok” ou se está “em alarme” (caso no qual as condições definidas são detectadas como a situação atual do sistema).

O parâmetro “ações” permite que ações sejam realizadas quando as condições forem satisfeitas, como enviar um email de alerta para determinados endereços de email.

1. Excesso de queries completadas por segundo:

- **Condições:** QueriesCompletedPerSecond > 0.15 para 1 pontos de dados em 1 dia
- **Utilidade:** Este alarme é útil para monitorar a carga de trabalho do sistema em relação ao número de consultas (queries) que está processando por segundo. Pode ajudar a identificar picos de tráfego inesperados ou avaliar se sua infraestrutura está lidando adequadamente com a demanda.

1. Tempo excedente de execução:

- **Condições:** QueryRuntimeBreakdown > 3880 para 1 pontos de dados em 5 minutos
- **Utilidade:** Este alarme é configurado para alertar quando o tempo de execução de uma operação específica excede um limite definido. Por exemplo, se você tem uma função serverless que normalmente leva apenas alguns segundos para ser concluída, um tempo de execução significativamente maior pode indicar problemas de desempenho ou gargalos.

1. Duração excessiva das queries

- **Condições:** QueryDuration > 1970000 para 1 pontos de dados em 5 minutos
- **Utilidade:** Este alarme monitora o tempo que as queries levam para serem executadas. Se o tempo de execução de uma consulta exceder a condição definida, o alarme será acionado. Isso é útil para identificar consultas que podem estar afetando o desempenho geral do sistema.

9. Análise de Impacto Ético

Em um mundo cada vez mais impulsionado por dados e tomada de decisão com base nos mesmos, a ética tornou-se um pilar fundamental para garantir que essas inovações promovam benefícios sociais, respeitem os direitos individuais e evitem consequências prejudiciais. A crescente intersecção entre tecnologia e dados requer uma abordagem cuidadosa e responsável, destacando a importância da análise de impacto ético. Esta documentação tem como objetivo fornecer um guia abrangente e sistemático para avaliar os impactos éticos associados a projetos, processos ou tecnologias que envolvem a manipulação e interpretação de dados. Ao abordar princípios fundamentais e estratégias práticas, esta documentação visa, também, a exploração de requisitos importantes para toda curadoria e utilização coerente das fontes de dados, com uma visão ética, assegurando que inovações tecnológicas sejam guiadas por princípios éticos e promovam um impacto positivo em todas as esferas da sociedade.

1. Privacidade e proteção de dados

Sobre os dados coletados, vale ressaltar que uma parte deles é de natureza pública, enquanto outra parcela é fornecida diretamente pelo cliente. No caso desta última, a coleta é realizada de forma segura por meio de uma API dedicada, assegurando o tráfego protegido dessas informações até a fase de armazenamento na nuvem. Para tal finalidade, adotamos uma abordagem robusta que envolve a utilização do Data Lake da AWS, mais especificamente o Amazon S3, em conjunto com um Data Warehouse, o Amazon Redshift. É relevante salientar que, no caso específico do Amazon S3, empregamos a criptografia padrão oferecida pela plataforma, assegurando assim a proteção dos dados armazenados desde o momento da coleta até a persistência na nuvem. Essa arquiteturameticulosamente planejada não apenas facilita a eficácia operacional, mas também assegura a integridade e confidencialidade dos dados em cada etapa do processo, garantindo, assim, a segurança abrangente do sistema.

É essencial garantir que as organizações responsáveis pelos dados utilizados no projeto trabalham com informações que não são diretamente ligadas a indivíduos específicos. Isso inclui

dados sobre a população, sobre o ambiente econômico e sobre o poder aquisitivo de uma região. É preciso validar se essas informações são coletadas e utilizadas para fins de pesquisa e análise, não para identificar ou rastrear indivíduos específicos.

Para isso, é necessário que se realize uma análise de cada fonte utilizada. Essas análises serão feitas considerando pontos como: se as informações podem ser consideradas sensíveis, ou seja, se são informações como dados de saúde, informações financeiras pessoais ou detalhes sobre comportamento pessoal; se foi garantido que a fonte está de acordo com o LGPD, o que pode ser feito checando por exemplo se os titulares dos dados têm ou não o direito de solicitar acesso aos mesmos; entre outras.

A ANVISA possibilita a análise de poder aquisitivo por região, uma vez que exerce a vigilância sanitária sobre produtos e serviços. Esta fonte apresenta dados de cunho econômico que não se caracterizam, portanto, como dados pessoais. O que garante a proteção dos dados é a chamada “Política de Proteção de Dados Pessoais” que seguem. A política visa, de acordo com o Ministério da Saúde, “garantir o cumprimento das normas relacionadas à privacidade, à transparência, ao acesso às informações públicas e à proteção das liberdades e dos direitos fundamentais dos indivíduos”.

O CNPJ armazena informações cadastrais das pessoas jurídicas e de outras entidades. O CNPJ em si é um dado não pessoal, pois ele é um número único que identifica uma entidade. Porém, as informações que podem ser obtidas a partir dele podem não ser. Por exemplo, a situação cadastral do CNPJ pode indicar se a empresa está em conformidade com suas obrigações fiscais e tributárias.

Apesar disso, ele pode ser classificado como uma fonte de dados não pessoais pois contém informações sobre a situação cadastral da empresa, que são relacionadas ao negócio da empresa, não a indivíduos. Ele não contém informações que possam ser usadas para identificar uma pessoa individualmente e, por fim, não contém informações que possam ser consideradas sensíveis.

O BACEN permite acesso a informações sobre taxa selic, inflação e o valor da moeda em dólar. Estes dados são de cunho estritamente econômico, não representando, portanto, ameaça a privacidade de terceiros.

O POF e o IBGE têm acesso a informações relacionadas ao poder aquisitivo de cada região, como a composição dos orçamentos domésticos e as condições de vida da população brasileira. Os dados que o POF fornece não são utilizados para análise de indivíduos específicos, portanto pode ser considerada uma fonte de dados não pessoais, assim como aquelas utilizadas do IBGE.

2. Equidade e justiça

Ao trabalhar com grandes volumes de dados, é necessário considerar os possíveis impactos em grupos específicos e buscar formas de minimizar as disparidades. O uso ético de dados em arquiteturas de big data envolve garantir que todos os grupos se beneficiem do uso desses dados e que ninguém seja prejudicado ou desfavorecido pelo uso indevido ou enviesado desses dados.

Por exemplo, o uso de dados do IBGE pode revelar disparidades socioeconômicas entre diferentes grupos, que podem ser amplificadas se as análises geradas não forem feitas de forma justa e equitativa. Da mesma forma, os dados da ANVISA podem ter implicações significativas para a saúde pública e podem impactar desproporcionalmente grupos vulneráveis se não forem manuseados de forma ética (De Freitas Saldanha et al., 2021).

Para minimizar as disparidades no uso de Big Data, é vital garantir que o pré-processamento dos dados seja realizado de maneira justa e equitativa . Isso pode envolver a garantia de que todos os grupos estejam adequadamente representados nos dados, a consideração de vieses potenciais nos dados e a implementação de técnicas de análise robustas para minimizar a chance de resultados tendenciosos. Garantir também que dados que podem ser relevantes para alguma minoria, por exemplo, não sejam removidos.

A visualização dos dados, a partir do infográfico, também tem que ser feita pensando nos princípios de ética, uma vez que a distorção de dados, mesmo que seja feita de forma acidental, pode acabar apoiando alguma narrativa falsa e intensificando uma possível desinformação. Adotar uma abordagem ética implica em evitar manipulações que possam distorcer a compreensão dos

dados, selecionar adequadamente os métodos de visualização para representar fielmente as informações e fornecer contexto apropriado para evitar interpretações equivocadas.

Além disso, é crucial considerar o impacto potencial nas comunidades marginalizadas ao lidar com dados sensíveis. O uso indevido dessas informações pode resultar em discriminação e reforçar desigualdades existentes. Para evitar isso, é imperativo implementar medidas rigorosas de segurança e privacidade, como a anonimização eficaz dos dados, para garantir que a identidade de grupos específicos não seja comprometida.

Em resumo, ao lidar com dados provenientes de fontes como BACEN, CNPJ, POF, IBGE, ANVISA, é fundamental adotar uma abordagem ética em todas as fases do processo, desde a coleta até a visualização, considerando cuidadosamente os impactos potenciais em grupos específicos e implementando medidas para minimizar disparidades e proteger a privacidade.

3. Transparência e consentimento informado

Primeiramente é importante ressaltar que o LGPD estabelece diretrizes sobre a coleta e o tratamento de dados pessoais, assegurando que o consentimento seja obtido. Todas as partes envolvidas, seja na pesquisa do IBGE ou no fornecimento de dados para o CNPJ, devem operar em conformidade com essas regulamentações.

O IBGE tem como princípio a transparência dos dados coletados. Antes de realizar qualquer pesquisa, o IBGE informa os objetivos das informações obtidas, garantindo que todos os participantes estejam cientes do propósito da coleta. No caso da POF, o IBGE adota medidas rigorosas para assegurar que os participantes estejam cientes da natureza da pesquisa, dos benefícios sociais associados e dos métodos de coleta de dados. Além disso, é assegurado que

os dados individuais sejam tratados com sigilo e confidencialidade, sendo utilizados apenas para fins estatísticos.

Quanto ao CNPJ, a obtenção de consentimento é uma parte fundamental do processo. As empresas e organizações que fornecem informações para o CNPJ são informadas sobre a necessidade e o propósito da coleta desses dados. O acesso a essas informações é estritamente regulamentado e destinado a fins específicos, como fiscalização e transparência.

A API do cliente disponibiliza um CSV com todas as vendas realizadas por CNPJ, apresentando a data, o valor e a quantidade atrelados a cada um. A este CSV se aplicam, portanto, as mesmas especificações citadas sobre as pesquisas do CNPJ.

Por fim, os dados coletados da ANVISA e do BACEN correspondem de maneira geral a dados econômicos não pessoais, situação na qual a transparência e o consentimento informado são garantidos por conta da natureza destes dados.

4. Responsabilidade social

No cenário dinâmico do avanço do big data, a responsabilidade social torna-se um elemento central que orienta o impacto desta tecnologia na sociedade. À medida que os dados se tornam cada vez mais abundantes, a adoção de práticas éticas e socialmente responsáveis torna-se não só mais um requisito, mas também crítica para mitigar as desigualdades, promover a inclusão e garantir que o desenvolvimento tecnológico tenha um impacto que beneficie a todos.

Ao analisar o impacto social dos projetos, nos comprometemos com uma avaliação criteriosa, focando em como essas iniciativas repercutem nas comunidades e no meio ambiente. O compromisso do projeto com as comunidades e meio ambiente é o mais importante para o grupo, principalmente para seguir as normas de regras da ODS. Além disso, não só enfatizando os impactos positivos esperados, mas também implementando medidas preventivas para evitar quaisquer impactos negativos.

O foco principal é garantir que os nossos esforços contribuam de forma eficaz e positiva para questões globais prementes, como a redução da desigualdade, o combate às alterações climáticas e a promoção da inovação tecnológica responsável.

Esta abordagem reflete o nosso forte compromisso não só em impulsionar a inovação e a eficiência, mas também em promover o progresso social e ambiental de uma forma ética e sustentável. Trabalhamos para integrar práticas sustentáveis e éticas em todas as fases dos projetos, traçando um caminho que não só antecipa as necessidades atuais, mas também visa construir um futuro mais equitativo e sustentável.

Concluindo, com o avanço do big data, a responsabilidade social tornou-se crucial. Ao alinhar os projetos de dados com os ODS, priorizamos não apenas a inovação e a eficiência, mas também o impacto positivo nas questões globais. A nossa abordagem visa não só antecipar as necessidades atuais, mas também construir um futuro mais equitativo e sustentável, integrando práticas éticas em todas as fases dos projetos. Desta forma, não só impulsionamos o progresso tecnológico, mas também promovemos o progresso social e ambiental ético e duradouro.

5. Viés e discriminação

O viés é um tópico importante quando tratamos sobre dados, especialmente quando considerando os mesmos na temática de ciência de dados e modelos de predição ou classificação. Neste sentido, o viés se baseia na distorção nos dados, ou tendência desvirtuada ou preconceituosa em relação ao conhecimento e insights extraídos após análise dos dados.

Esses vieses podem ser definidos em três principais categorias, o viés de amostragem, o viés humano e viés algorítmico (Luis B, 2023). Respectivamente, o primeiro se refere aos tipos de dados provenientes de amostras bem pequenas (pode ser até pelo nicho de pesquisa performado), que não representam a população no geral. O viés humano, é um dos vieses mais recorrentes quanto estamos no processo de tratamento, análise e interpretação dos dados. Esse viés se dá especialmente pela opinião, possíveis preconceitos e estereótipos que possam guiar a interpretação e inferência dos dados, assim como influenciar nas decisões e resultados finais. Por fim, o viés algorítmico se reflete ao processo de aprendizado de máquina, principalmente em modelos de aprendizado supervisionado, ser feito a partir de conjuntos de dados refletem

desigualdades sociais, discriminação ou estereótipos existentes (isso acontece muito com dados antigos).

Sendo assim, é muito importante que tenhamos isso em mente, sempre que formos tratar com dados, especialmente, aqueles no qual não houve uma curadoria na forma em que foi feita a pesquisa, em base de dados estatísticos e dados públicos (dados abertos de livre acesso). Principalmente, tendo em vista que esses vieses podem surgir de forma intencional ou não intencional.

Com base em nossas fontes de dados, podemos examinar as potenciais fontes de viés e estratégias para mitigá-las em cada frente:

1. Dados da ANVISA:

- Possível Viés: Pode haver viés socioeconômico nas petições alimentícias, favorecendo áreas com maior poder aquisitivo.
- Mitigação: Normalizar os dados em relação à população de cada região, considerar indicadores socioeconômicos locais e aplicar técnicas estatísticas para ajuste de viés.

2. Dados do Bacen:

- Possível Viés: Dados financeiros podem refletir desigualdades econômicas e impactar a análise de diferentes regiões.
- Mitigação: Normalizar os dados em relação à população, considerar fatores socioeconômicos, e adotar técnicas de correção para equilibrar possíveis distorções.

3. Dados de CNPJ:

- Possível Viés: Pode haver desigualdades na representação de diferentes canais de venda, favorecendo alguns em detrimento de outros.
- Mitigação: Analisar representatividade proporcional dos canais de venda, considerar dados adicionais sobre distribuição demográfica e aplicar técnicas de ajuste.

4. Dados da POF e IBGE:

- Possível Viés: As amostras podem não ser totalmente representativas, resultando em visões distorcidas das condições de vida.

- Mitigação: Validar a representatividade das amostras, corrigir distorções conhecidas e incorporar dados adicionais para enriquecer a análise.

5. Dados coletados com a API do parceiro:

- Possível Viés: Pode haver distorções nas dinâmicas de vendas fictícias que não refletem a realidade do mercado.
- Mitigação: Validar a consistência dos dados, considerar fatores contextuais que possam influenciar as vendas fictícias e aplicar técnicas de ajuste.

Na estratégia geral de mitigação, é fundamental adotar medidas que assegurem a integridade e imparcialidade das análises de dados. A transparência é importante, exigindo que os métodos de coleta, processamento e análise sejam comprehensíveis e acessíveis a todos os envolvidos. O segundo ponto principal é a revisão ética, realizada de forma regular, para identificar potenciais pontos de discriminação ou exclusão involuntária, garantindo que os resultados sejam éticos e equitativos. Além disso, é imperativo fomentar a diversidade na equipe responsável pela análise, tendo em vista que uma equipe com perspectivas variadas contribui para uma abordagem mais abrangente e sensível às diferentes nuances presentes nos dados. Consequentemente, a inclusão de diversas vozes na tomada de decisões promove uma análise mais completa e objetiva. A reavaliação contínua é outra peça-chave na estratégia de mitigação. Estabelecer protocolos que permitam revisar e ajustar as análises à medida que novos insights ou preocupações éticas surgem é fundamental. Esse processo dinâmico assegura que as análises permaneçam alinhadas com os padrões éticos mais recentes e com a evolução do contexto em que estão inseridas. Em conjunto, essas abordagens formam uma base sólida para a realização de análises de dados éticas e equitativas, mitigando potenciais riscos de viés e discriminação.

Em conclusão, a análise de dados enfrenta desafios significativos relacionados ao viés, especialmente nas áreas da ciência de dados e modelos de predição. O viés, caracterizado pela distorção nos dados e tendências preconceituosas, pode manifestar-se em diferentes formas, como o viés de amostragem, o viés humano e o viés algorítmico. Cada uma dessas categorias apresenta riscos específicos que podem comprometer a precisão e a equidade das análises.

Ao examinar as fontes de dados específicas, identificamos possíveis fontes de viés em cada frente. Contudo, é crucial adotar uma abordagem proativa para mitigar esses vieses e garantir a integridade das análises. A estratégia geral de mitigação destaca a importância da transparência nos métodos de coleta e análise, da revisão ética regular para identificar possíveis discriminações, da promoção da diversidade na equipe para trazer perspectivas variadas e da reavaliação contínua para ajustar as análises conforme necessário.

Essas medidas, quando implementadas de forma coordenada, formam uma base robusta para a realização de análises éticas e equitativas, minimizando os riscos associados ao viés e à discriminação nos dados. Assim, ao enfrentar os desafios inerentes à análise de dados, é essencial manter um compromisso contínuo com a transparência, a ética e a diversidade para garantir resultados embasados e sem, ou com pouco viés.

10. Pipeline Integrado

10.1 Descrição

Este projeto consiste em um pipeline de dados completo, integrando os processos de Extração, Transformação e Carga (ETL) para dados de CNPJ, POF, BACEN, DENSIDADE e ANVISA. Além disso, inclui a implementação de um modelo de regressão utilizando o algoritmo RandomForestRegressor e a geração de um gráfico de importância de recursos.

10.2 Configuração e Execução

Pré-requisitos

Antes de executar este projeto, você precisa instalar as dependências necessárias. As dependências estão listadas no arquivo requirements.txt. Para instalá-las, execute o seguinte comando no terminal:

```
pip install -r requirements.txt
```

Configuração

Você precisa fornecer suas credenciais da AWS e do Redshift para que o script possa carregar os dados transformados no seu bucket S3 e se conectar ao banco de dados, respectivamente. Substitua 'aws_access_key_id', 'aws_secret_access_key', 'bucket_name', 'aws_session_token', 'host', 'port', 'user', 'password' e 'database' pelos seus valores reais no código.

Execução

Para executar o pipeline completo, você precisará executar três scripts em sequência:

`api_etl.py` - Este script executa o ETL para dados de vendas. Ele busca dados de uma API, transforma os dados e, em seguida, os carrega em um bucket AWS S3.

`base_etl.py` - Este script processa arquivos CNPJ, POF, BACEN, DENSIDADE(IBGE) e ANVISA. Portanto, você precisará modificar o caminho do arquivo no código para cada arquivo que deseja processar e executar o script novamente.

`modelo.py` - Este script executa um modelo ensemble usando o algoritmo RandomForestRegressor. Ele se conecta a um banco de dados Redshift, executa uma consulta SQL, transforma o resultado em um DataFrame pandas e, em seguida, treina o modelo de regressão. Além disso, ele gera um gráfico de importância de recursos que é salvo localmente.

Para executar cada script, navegue até o diretório que contém o script e execute o seguinte comando:

```
python <nome_do_script>.py
```

Substitua <nome_do_script> pelo nome do script que você deseja executar (api_etl, base_etl ou modelo).

10.3.1 Modelo Ensemble no Final do Pipeline

Modelo de Regressão

O modelo de regressão utilizado é o RandomForestRegressor, um algoritmo de aprendizado de máquina que utiliza múltiplas árvores de decisão para realizar previsões. Este modelo é treinado com um conjunto de características (features) para prever a quantidade de um produto.

Features

As características utilizadas para treinar o modelo são:

'preco': o preço do produto

'idCategory': a categoria do produto

'sigla_uf': a sigla do estado onde o produto é vendido

'taxa_selic': a taxa Selic no momento da venda

'taxa_exp_inflacao': a taxa de inflação esperada no momento da venda

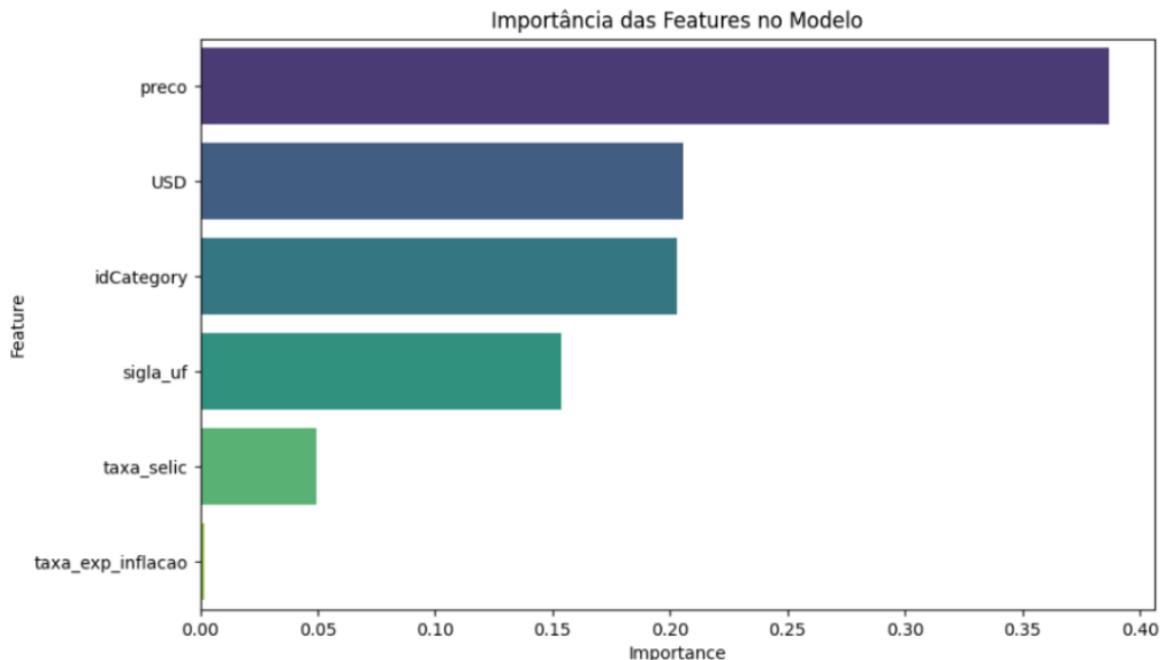
'USD': o valor do dólar no momento da venda

Query do Redshift

A query do Redshift é utilizada para obter os dados necessários para treinar o modelo. Ela seleciona as colunas 'produto', 'preco', 'quantidade', 'usd', 'taxa_selic', 'taxa_exp_inflacao', 'densidade_populacional_habitantes_km2', 'area_km2' e 'numer0_populacional' da tabela 'cnpj_vendas_bacen_final' e realiza um INNER JOIN com a tabela 'densidade_g05_real' na coluna 'id_municipio'.

10.3.2 Modelo Ensemble Infográfico

O script `modelo.py` gera um gráfico de importância de recursos que mostra a importância de cada recurso no modelo de regressão. Este gráfico é salvo localmente como um arquivo PNG. Você pode encontrar este arquivo no mesmo diretório do script. Por favor, note que cada vez que você executa o script, o gráfico de importância de recursos é atualizado.



11. Referências

Situação Cadastral CNPJ: entenda o que é e como funciona. (2023, December 7). Blog C6 Bank. <https://www.c6bank.com.br/blog/situacao-cadastral-cnpj>

Conheça a Política de Proteção de Dados Pessoais da Anvisa. (2023, October 23).

Agência Nacional De Vigilância Sanitária - Anvisa.

<https://www.gov.br/anvisa/pt-br/assuntos/noticias-anvisa/2023/conheca-a-politica-de-protecao-de-dados-pessoais-da-anvisa>

De Freitas Saldanha, R., Barcellos, C., & De Moraes Pedroso, M. (2021). Ciência de dados e big data: o que isso significa para estudos populacionais e da saúde? *Cadernos Saúde Coletiva*, 29(spe), 51–58.

<https://doi.org/10.1590/1414-462x202199010305>

Totvs, E. (2023, May 3). *Big Data: o que é, como funciona e como aplicar?* TOTVS.

<https://www.totvs.com/blog/inovacoes/big-data>

B, L. (2023, July 24). *O viés na Ciência de Dados.*

<https://www.linkedin.com/pulse/o-vies-na-ciencia-de-dados-luis-balero/?trackingId=>

a

10. Referências

Transformações digitais no Brasil: insights sobre o nível de maturidade digital das empresas no país. (1 C.E., January 1). McKinsey & Company.

<https://www.mckinsey.com/br/our-insights/transformacoes-digitais-no-brasil>

Negócio, R. V. M. (2023, February 23). As perspectivas e tendências para o varejo alimentar em 2023. *Vivo Meu Negócio*.

<https://vivomeunegocio.com.br/bares-e-restaurantes/gerenciar/varejo-alimentar/>

Statista. (n.d.). *Business Intelligence Software - Brazil | Market forecast*.

<https://www.statista.com/outlook/tmo/software/enterprise-software/business-intelligence-software/brazil>

Alimentos (n.d.). Investe SP.

[https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca%20de%2028%25%20da,e%20Estatística%20\(IBGE\)%20-%20](https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca%20de%2028%25%20da,e%20Estatística%20(IBGE)%20-%20)

Value Proposition Canvas: o que é e como funciona essa metodologia? - G4 Educacão. (n.d.). <https://g4educacao.com/portal/value-proposition-canvas>

Exemplos e práticas recomendadas de arquitetura de referência. (n.d.). Amazon Web Services, Inc.

https://aws.amazon.com/pt/architecture/?cards-all.sort-by=item.additionalFields.sortDate&cards-all.sort-order=desc&awsf.content-type=*all&awsf.methodology=*all&awsf.tech-category=*all&awsf.industries=*all&awsf.business-category=*all

O que é pipeline de dados? - Explicação sobre pipeline de dados - AWS. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/pt/what-is/data-pipeline/>

Monitoramento de aplicações e infraestrutura – Amazon CloudWatch – Amazon Web Services. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/pt/cloudwatch/>

Amazon Redshift – Data Warehouse na nuvem – Amazon Web Services. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/pt/redshift/>

Computação sem servidor - AWS Lambda - Amazon Web Services. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/pt/lambda/features/>

What is AWS Glue? - AWS Glue. (n.d.).

<https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html>

De Souza, I. (2021, February 12). *O que é TLS e quais são as diferenças entre ele e SSL? Descubra agora.* Rock Content - BR. <https://rockcontent.com/br/blog/tls/>

AWS IAM - Identity and Access Management - Amazon Web Services. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/pt/iam/>

O que é o IAM? - AWS Identity and Access Management. (n.d.).

https://docs.aws.amazon.com/pt_br/IAM/latest/UserGuide/introduction.html

Ka, N. (2021, December 16). *A Guide on Designing AWS Data Architectures* - narjes ka - Medium. *Medium.*

<https://medium.com/@karmeni.narjes.pro/a-guide-on-designing-aws-data-architectures-6a488ef9260c>

Hellmuth, M. (2023, March 10). *10 Best practices for creating Effective wireframes.* Medium.

<https://medium.com/design-with-figma/10-best-practices-for-creating-effective-wireframes-a7e1dc94125e>

Practical Tips for creating Better Wireframes | Wireframing Academy | Balsamiq. (n.d.).

<https://balsamiq.com/learn/articles/practical-tips-for-better-wireframes/>

Rodríguez, A. (2022, November 21). *O que são Heurísticas de Nielsen e como aplicá-las em UX*. Rock Content - BR. <https://rockcontent.com/br/blog/heuristicas-de-nielsen/>

Material didático - IMD. (n.d.). <https://materialpublic.imd.ufrn.br/curso/disciplina/3/73/14/2>

User. (2023, April 12). *5 benefícios em usar o Redshift para Data Warehouse*. eMaster Cloud E Security.

<https://emaster.cloud/data-analytics/5-beneficios-em-usar-o-redshift-para-data-warehouse/>

Material didático - IMD. (n.d.). <https://materialpublic.imd.ufrn.br/curso/disciplina/3/73/14/2>

User. (2023, April 12). *5 benefícios em usar o Redshift para Data Warehouse*. eMaster Cloud E Security.

<https://emaster.cloud/data-analytics/5-beneficios-em-usar-o-redshift-para-data-warehouse/>

Usar alarmes do Amazon CloudWatch - Amazon CloudWatch. (n.d.).

https://docs.aws.amazon.com/pt_br/AmazonCloudWatch/latest/monitoring/AlarmThatSendsEmail.html

. Anexos