



Integration

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, conseqüentemente, 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

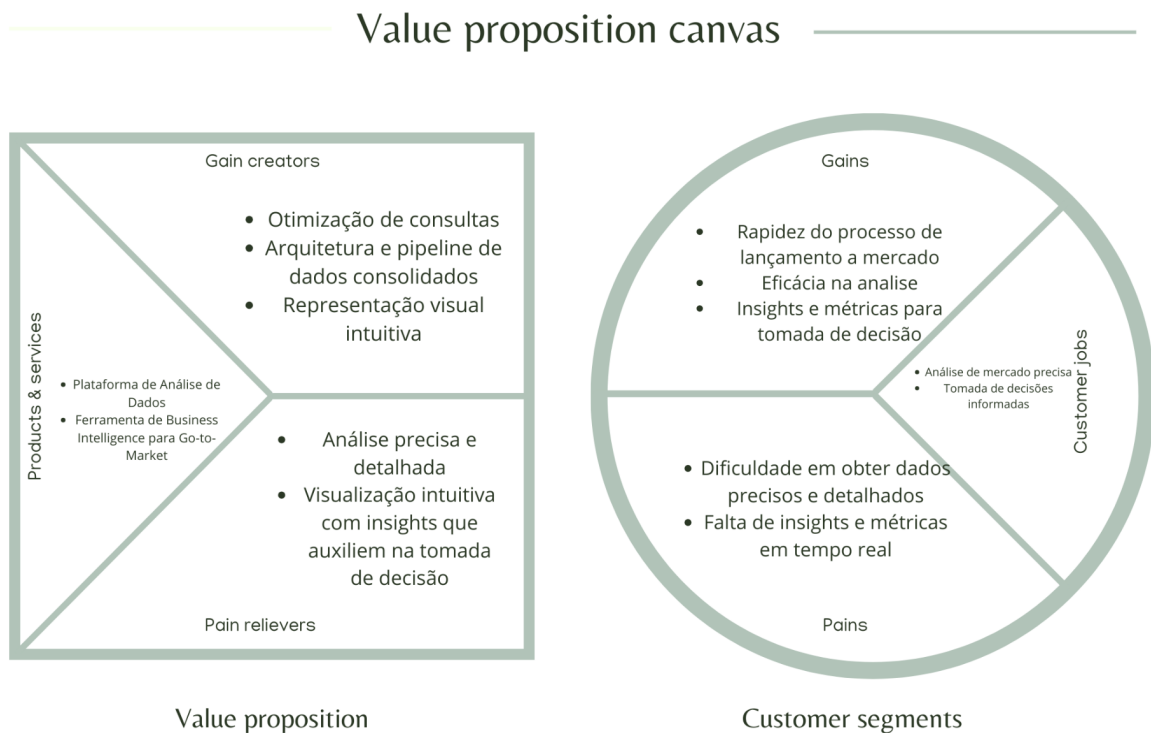
O problema em questão envolve um cliente distribuidor, da Integration, que opera em diversas categorias nos setores alimentar e de food service. Esse cliente busca uma ferramenta que o auxilie a entender o potencial de consumo de cada categoria em níveis detalhados, como cidade e canal de atendimento 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				
90%	Ameaça 9: Dependência de terceiros, como fornecedores da nuvem	-	-	-	Ameaça 2: Limitação em integrações multclouds, e serviços pouco escaláveis em alguns lugares	-	-	Oportunidade 3: Desenvolvimento de infográficos, com dados em tempo real, e previsão analítica do consumo nas categorias pré-definidas	-	-
70%	-	-	Ameaça 8: Vazamento de dados ou chave de acesso por parte dos grupos	-	-	Oportunidade 1: Criação de ferramenta que auxilie na tomada de decisão dos cliente	Oportunidade 2: Realização de testes que comprovem uma alta precisão da ferramenta	-	-	-
50%	Ameaça 6: Dados imprecisos, incompletos ou de baixa qualidade podem prejudicar a análise estatística.	Ameaça 7: O cliente pode não adotar ou usar efetivamente o infográfico gerado.	-	-	Ameaça 1: Mau tratamento dos dados, causando ingestão de dados com baixa volumetria	-	-	Oportunidade 4: Criação de um pipeline de dados bem estruturado	-	-
30%	-	Ameaça 3: Alta expectativa do cliente e abstração do escopo do projeto	Ameaça 5: Baixa granularidade e insights enviesados (sem embasamento do cliente) na análise estatística	-	Ameaça 4: Criação de infográficos que não auxiliam na tomada de decisão ao cliente	-	-	Oportunidade 5: Redução de custo no uso de uma infraestrutura cloud linear, separando serviços em diferentes clouds	-	-
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: Mau tratamento dos dados, causando ingestão de dados com baixa volumetria.

- **Plano de Ação:** Realizar uma avaliação da infraestrutura atual identificando pontos fracos e gargalos, definindo o que pode ser feito.
- **Responsável:** Eric

- **Justificativa:** O risco de mau tratamento dos dados, resultando em baixa volumetria, é crítico devido ao potencial comprometimento da qualidade das informações.
- **Risco:** 50%

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: Vazamento de dados ou chave de acesso por parte dos grupos.

- **Plano de Ação:** Validar os "commits" (principalmente na main) e validar com o cliente se há necessidade de mascarar dados ou realizar anonimização.
- **Responsável:** Rodrigo Martins
- **Justificativa:** Quando dados ou chave de acesso são vazados, dá a oportunidade pros concorrentes ou Hackers invadirem o sistema e roubar todos os dados.
- **Risco:** 70%

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

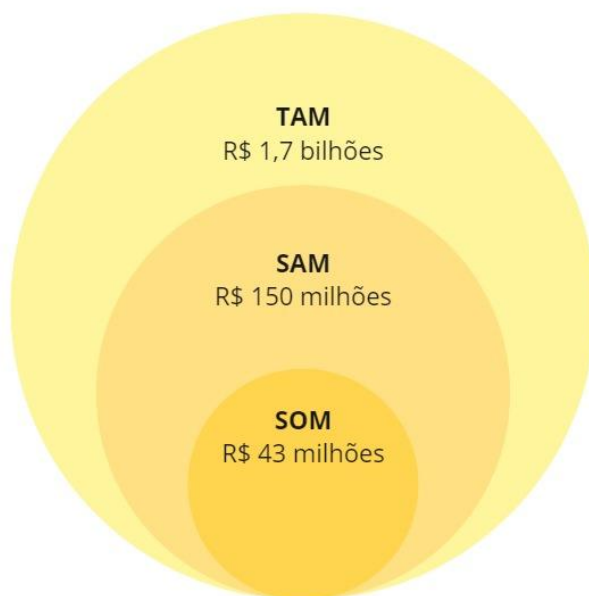
- **Plano de Ação:** Validar os "commits" (principalmente na main) e validar com o cliente se há necessidade de mascarar dados ou realizar anonimização.

- **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 compreensíveis para que haja conhecimento nos dados inicialmente coletados. Ricardo está comprometido em fornecer insights acionáveis e contribuir para o sucesso do projeto em questão.

RICARDO SILVA

CIENTISTA DE DADOS



NOME: Ricardo Silva

IDADE: 32 anos

GÊNERO: Masculino

LOCALIZAÇÃO: São Paulo, Brasil

RENDA: Por volta de R\$ 9 mil

OCUPAÇÃO: Cientista de dados

BIOGRAFIA

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 compreensíveis para que haja conhecimento nos dados inicialmente coletados. Ricardo está comprometido em fornecer insights acionáveis e contribuir para o sucesso do projeto em questão.

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 gargalos



“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

RENDIA: 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

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:

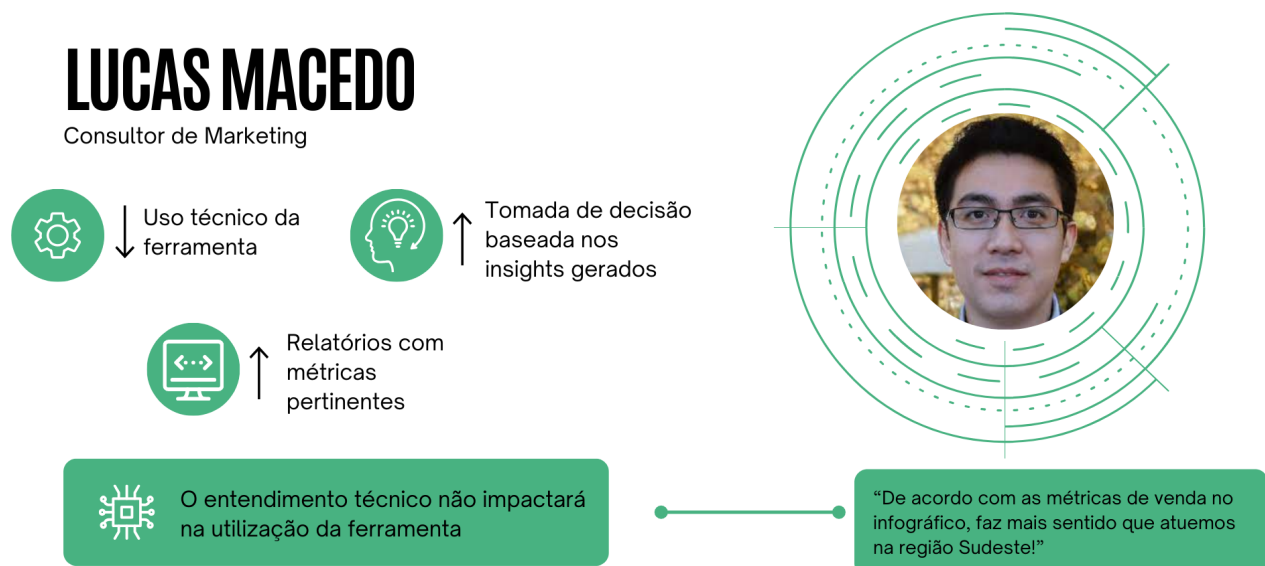
HÁBITOS RELACIONADOS AO SISTEMA:

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

- 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>"Preciso acessar os dados para diagnosticar o problema do cliente da melhor maneira possível."</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>"Esta solução parece a mais adequada para atender às nossas necessidades, mas não tenho certeza se os resultados serão satisfatórios"</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>"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"</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>"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."</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: <i>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.</i>
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 <u>Aceito:</u> <i>A infraestrutura está pronta para processar e armazenar os dados</i>

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 vem de forma não estruturados, com diferentes formatos de arquivos, desde CSV até arquivos em texto.

5.1.2 Fontes de Dados:

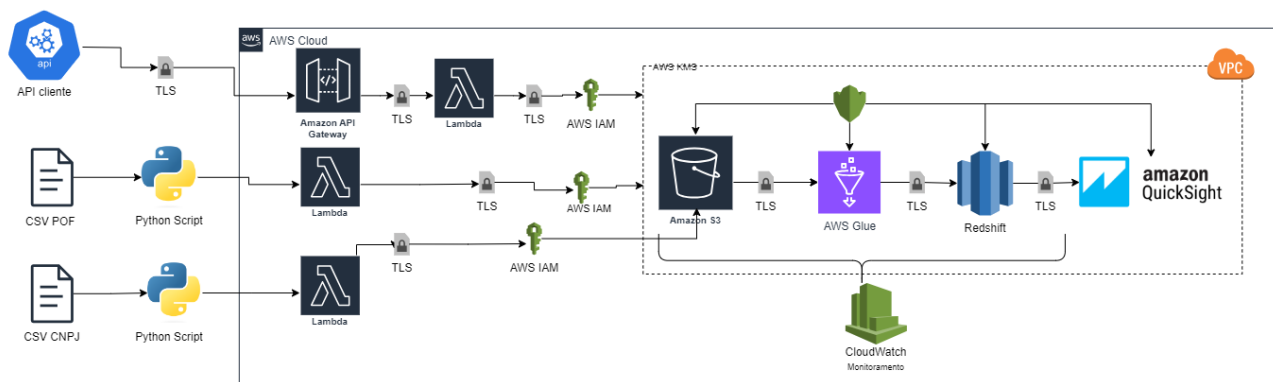
- Dados de censos/pesquisas do governo (CSV)
- Dados de CNPJs (CSV)
- Dados fornecidos pelo parceiro (API)

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 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.

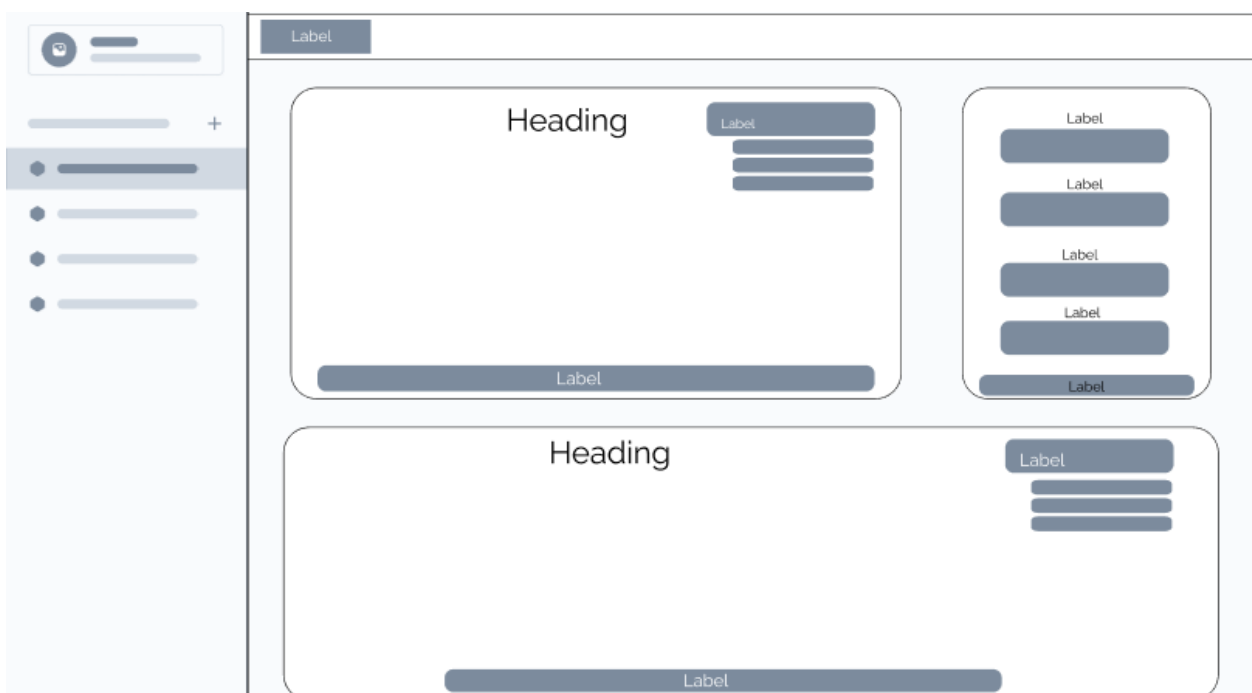
Para complementar a segurança, vamos fazer uso do AWS Key Management Service (KMS). O KMS pode ser integrado aos serviços que planejamos utilizar e nos permite criar Customer Master Keys (CMKs), que são essenciais para criptografar e decifrar os dados de forma segura. Além disso, o KMS é responsável 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 cumprir 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 partido 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:

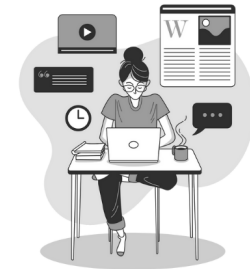
- Recomendação para sempre considerar a análise cruzada de dados de categoria e canal, proporcionando ao usuário uma visão mais abrangente das informações;
- Sugestão de incorporar um filtro de data, permitindo que o usuário selecione o período temporal desejado para a visualização;
- Dimensão geográfica: 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

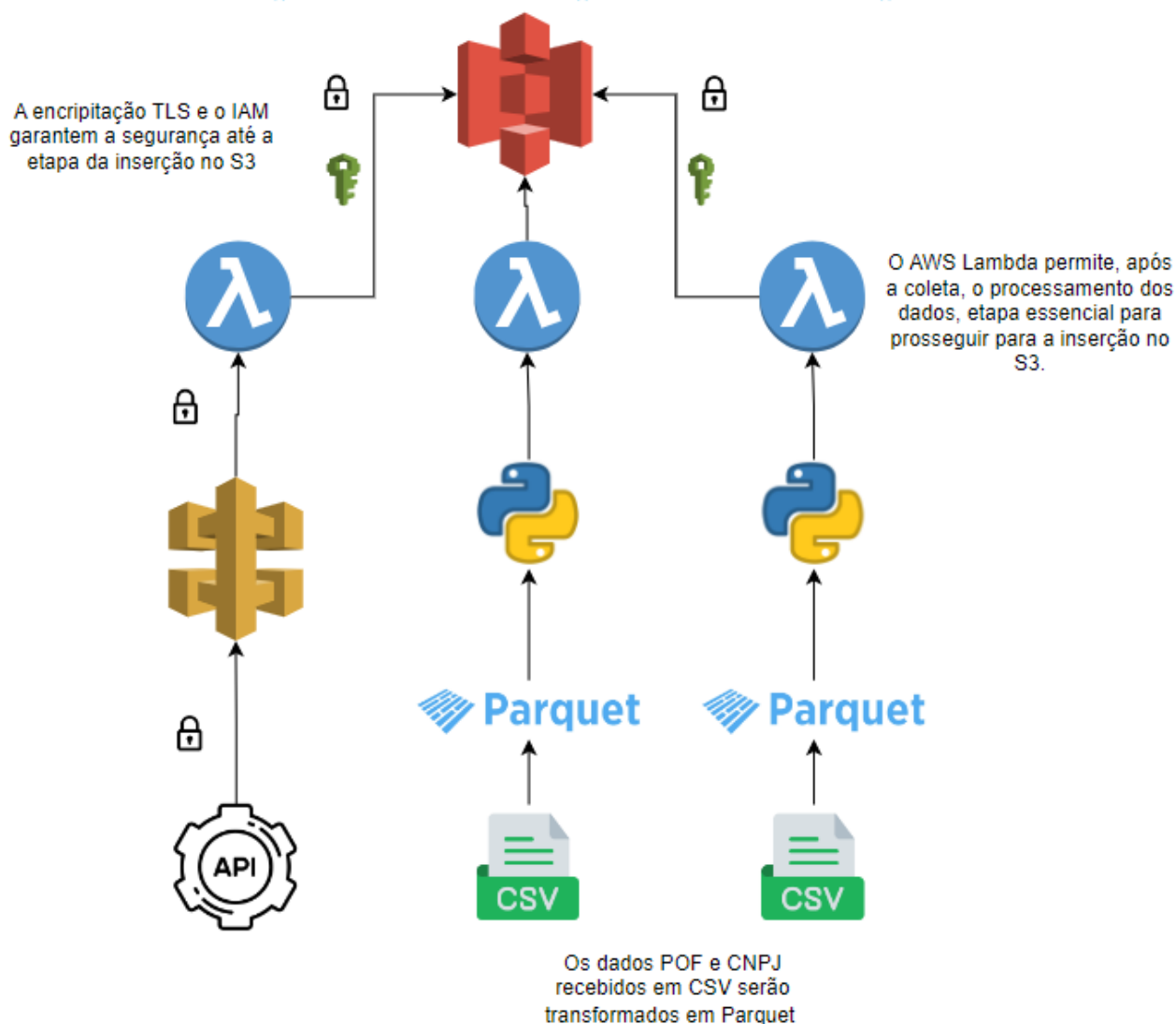
Nesta etapa da arquitetura, a Ingestão de Dados, compreende duas abordagens principais:

- 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

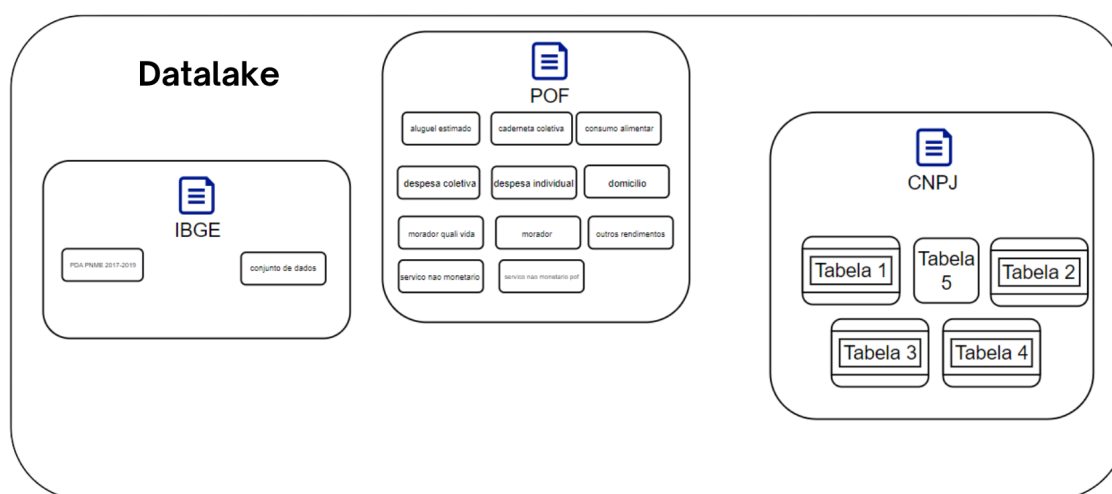
Para contextualização, o Datalake é basicamente um repositório que armazena grandes volumes de dados brutos, estruturados e não estruturados, em seu formato original. Diferente de um data warehouse, que segue um esquema de dados estruturado, o datalake permite realizar o armazenamento dos dados em sua 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

myawsbucket

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

Leste dos EUA (Norte da Virgínia) us-east-1

Copiar configurações do bucket existente - *opcional*

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

Escolher bucket

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.

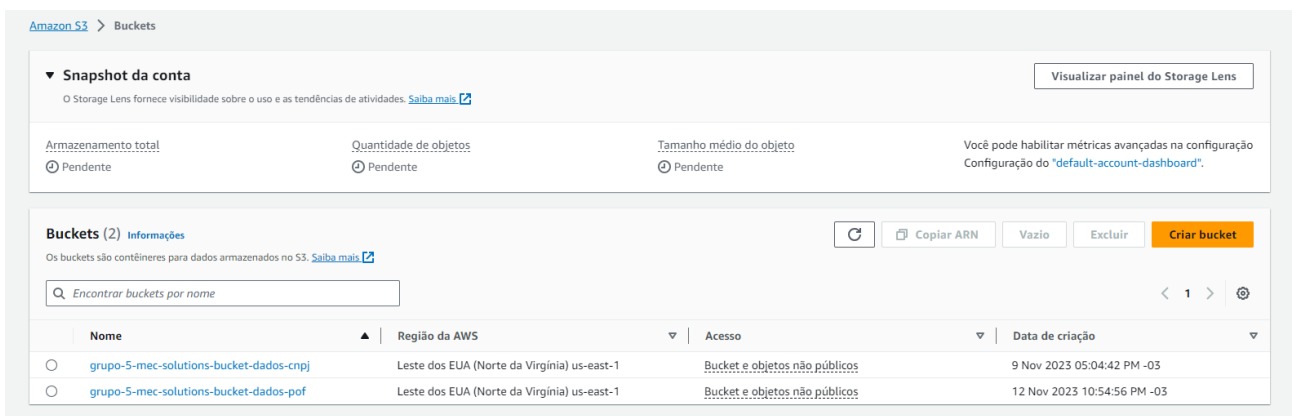
Criar bucket

Passo 8: Acesse o Bucket:

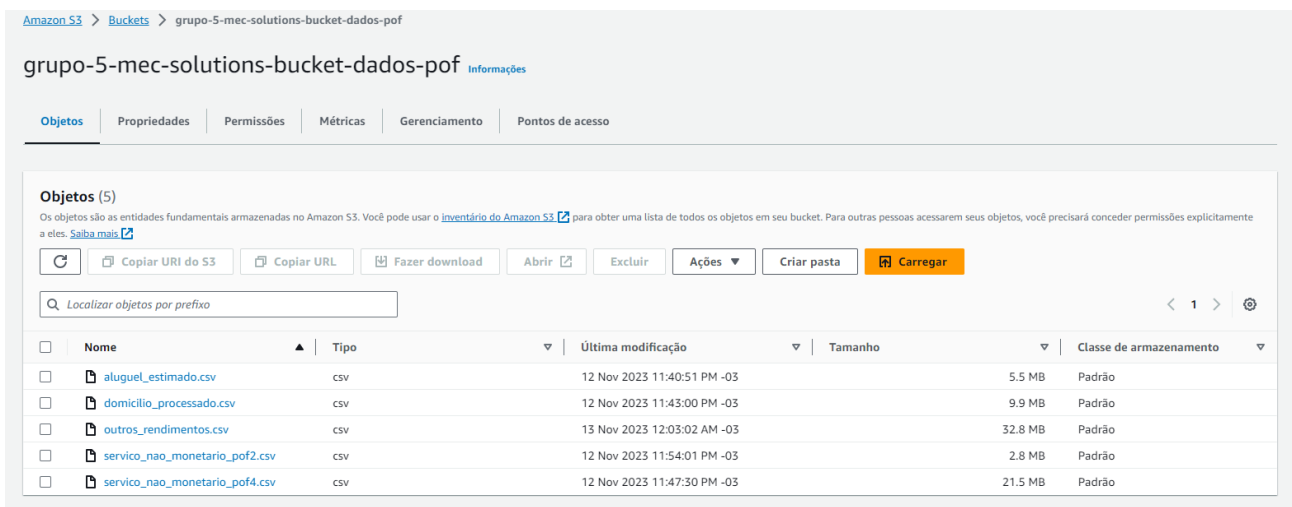
- Após a criação, você verá seu novo bucket listado no painel do Amazon S3. Clique nele para acessar e começar a gerenciar objetos (arquivos) dentro do bucket.

7.6 Dados alimentados no Datalake (exemplo CNPJ e POF)

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



Tabelas *tratadas e que subidas automaticamente no datalake:



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

Os próximos passos para a Sprint 2 serão a implementação do VPC, estruturação da ingestão de dados usando Lambda, Redshift (Data warehouse), S3 (Data Lake) e Glue (ETL) e configurações do API-Gate. Além disso, será realizada a prototipação da visualização dos dados.

9. Conclusões

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 de 28%2C6%25 da,e Estatística \(IBGE\) –](https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca de 28%2C6%25 da,e Estatística (IBGE) –)

Value Proposition Canvas: o que é e como funciona essa metodologia? - G4 Educação.

(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/>

. Anexos