



# **BIG DATA**

## **INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE**

# **INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI**

## **INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA**

### **INTEGRATION**

**Autores:**

Daniel Barzilai

Dayllan de Souza Alho

Gustavo Monteiro

Lucas de Britto Vieira

Luiz Augusto Pompeo de Camargo Franco Ferreira

Matheus Fidelis dos Santos Pinto

Pedro de Carvalho Rezende

**Data de criação:** 16 de Outubro de 2023

SÃO PAULO – SP

2023

# Sumário

<b>Sumário</b>	<b>3</b>
<b>1. Introdução</b>	<b>5</b>
1.1. Parceiro de Negócios	5
1.2. Definição do Problema	6
1.2.1. Contexto	6
1.2.2. Natureza do Problema	6
<b>2. Objetivos</b>	<b>7</b>
2.1. Objetivos Gerais	7
2.2. Objetivos Específicos	7
2.3. Justificativa	8
<b>3. Compreensão do Problema</b>	<b>8</b>
3.1 Canvas Proposta de Valor	8
3.2. Matriz de Risco	10
3.3. TAM SAM SOM	12
<b>4. Lean Inception</b>	<b>14</b>
4.1. O Produto (É – Não É – Faz – Não Faz)	14
4.1.1 O que o Sistema É:	15
4.1.2. O que o Sistema Não É:	15
4.1.3. O que o Sistema Faz:	16
4.1.4. O que o Sistema Não Faz:	16
<b>5. Análise de Experiência do Usuário</b>	<b>18</b>
5.1. Personas	18
5.2. Jornada do Usuário	20
5.3. User Stories	21
6.1. Instituto Brasileiro de Geografia e Estatística (IBGE)	36
6.2. Pesquisa de orçamento familiar (POF)	36
6.3. RAIS e CAGED Microdados	37
6.4. Receita Federal Dados Abertos	37
6.5. Dados Abertos MEC	37
6.6. Dados Abertos INEP	39
6.7. Open Data SUS	39
6.8. Conjunto de dados de Códigos Postais Mundiais	39
<b>7. Arquitetura da Solução</b>	<b>41</b>
7.1. Arquitetura Versão Preliminar	41
7.1.1 Identificação dos dados de entrada e saída	42
7.1.1.1. Dados de Entrada	42
7.1.1.2. Dados de Saída	42
7.1.2. Especificação de Entrada e Saída de cada módulo	42
7.2. Arquitetura Final	44
7.3. Detalhes Técnicos do Pipeline de dados na AWS	46

7.5. Modelo de Regressão para prever o coeficiente do nível de "Pobreza Multidimensional Não Monetária" (IPM-NM)	49
7.6. Infográfico - Relatório de Análise de Eficácia e Sugestões de Melhoria no Projeto	50
<b>8. Análise de Custos e Impacto da Solução</b>	<b>53</b>
8.1. Impactos Esperado da Solução	53
8.2. Análise de Custos	53
8.2.1. Serviços da Amazon Services Utilizados:	53
8.2.1.1. Amazon Redshift:	53
8.2.1.2. S3 (S3 Standard e Data Transfer):	54
8.2.1.3. AWS Lambda:	54
8.2.1.4. Amazon EC2 (Elastic Compute Cloud):	54
8.2.2. Custos Projetados para o Primeiro Ano:	54
8.2.2.1. Investimento Inicial:	54
8.2.2.2. Custo das ferramentas:	54
8.2.2.3. Custo Total:	55
8.2.3. Observação	56
<b>9. Prototipação em Baixa Fidelidade com Wireframes</b>	<b>57</b>
9.1. Wireframes Desenvolvidos	57
9.2. Técnicas Aplicadas	61
9.2.1. Ênfase na Prioridade dos Dados	62
9.2.2. Organização de Informações para Análise Cruzada	62
9.2.3. Dados comparativos por Rankeamento de Variáveis	62
9.2.4. Visualização Total com Opção de Filtros	62
9.2.5. Status dos Dados Utilizados	63
9.2.6. Telas Menores e Gestos Típicos de Dispositivos Móveis	63
9.2.7. Reduzir a Utilização de Texto:	63
9.2.8. Incorporação de Feedbacks:	64
<b>10. Análise de Impacto Ético</b>	<b>64</b>
10.1. Privacidade e Proteção de Dados	65
10.2. Equidade e Justiça	66
10.3. Transparência e Consentimento Informado	67
10.4. Responsabilidade Social	67
10.5. Viés e Discriminação	68
10.6. Discussão	69
<b>11. Plano de Comunicação</b>	<b>71</b>
11.1. Objetivo	71
11.2 Stakeholders	71
11.2.1. Grupo Integration Consulting	72
11.2.2. Grupo Inteli	72
11.2.3. Tipologia de stakeholders	72
11.2.4. Classificação dos stakeholders	74
11.2.4.1 Stakeholders Integration Consulting	74

11.2.4.2 Stakeholders Inteli	74
<b>11.3. Mensagens-chave</b>	<b>74</b>
11.3.1. Mensagem-Chave para Integration Consulting	75
11.3.2. Mensagem-Chave para Inteli	75
11.4. Canais de comunicação	75
11.5. Plano de implementação	76
11.6. Medidas de sucesso, Feedback e Ajustes	77
<b>12. Anexos</b>	<b>78</b>
12.1. Matriz de risco	78
<b>13. Conclusões</b>	<b>79</b>
<b>14. Referências</b>	<b>80</b>

---

## 1. Introdução

O projeto proposto visa à exploração e aplicação de práticas avançadas relacionadas ao Big Data, visando proporcionar uma compreensão abrangente desta área em constante crescimento. Através de uma parceria com a Integration Consulting, pretende-se desenvolver uma solução robusta para uma empresa distribuidora, utilizando conceitos avançados e infraestrutura fornecida pela AWS. O projeto aborda diversas etapas, desde a coleta inicial de dados até a aplicação de técnicas avançadas de análise. Serão explorados temas como ingestão de dados, pré-processamento em data lakes, mineração de dados, visualização de dados e análise automatizada, com foco em métodos preditivos, prescritivos e diagnósticos.

O uso da infraestrutura da AWS, incluindo serviços como S3, Redshift e AWS Lambda, permitirá o gerenciamento e análise eficiente de grandes volumes de dados, com o intuito de extrair insights estratégicos para determinada empresa. Além disso, a criação de um infográfico resumindo os resultados do estudo estatístico proporcionará uma entrega de valor mais eficiente e abrangente ao cliente.

Durante o desenvolvimento do projeto, será enfrentado desafios específicos relacionados à gestão de dados em ambientes empresariais e de mercado, com foco na escalabilidade e disponibilidade em ambientes distribuídos em nuvem. Esses desafios serão abordados por meio da aplicação prática de conceitos avançados em Big Data, buscando soluções que atendam às necessidades da empresa distribuidora.

## **1.1. Parceiro de Negócios**

A Integration Consulting é uma empresa especializada em consultoria e soluções de tecnologia. Com uma vasta experiência no mercado, eles oferecem serviços de alta qualidade para seus clientes, ajudando-os a alcançar seus objetivos de negócios por meio de soluções personalizadas e inovadoras. Com um compromisso com a excelência, a Integration Consulting é reconhecida por sua expertise em tecnologia e pela capacidade de oferecer resultados eficientes.

## **1.2. Definição do Problema**

O problema em questão envolve a necessidade de criar um sistema de análise de dados abrangente para compreender e explorar as informações contidas em diversas fontes de dados governamentais, dados de CNPJ e dados de parceiros. O objetivo principal é gerar um conjunto de informações valiosas que ajudarão a Integration Consulting e seus clientes a tomar decisões estratégicas informadas no setor de distribuição, com ênfase em consumo e vendas por região e ao longo do tempo.

### **1.2.1. Contexto**

A Integration lida com uma ampla variedade de informações relacionadas a seu cliente, um distribuidor atuante em várias categorias nos canais alimentar e de food service. Esses dados são obtidos de diversas fontes, incluindo dados governamentais, registros de CNPJ e informações de parceiros comerciais. A integração e análise eficientes dessas fontes de dados são essenciais para direcionar ações estratégicas e melhorar o desempenho do cliente.

### **1.2.2. Natureza do Problema**

O problema consiste em três desafios principais:

Integração de Dados: Integrar e processar dados de diferentes fontes, incluindo dados governamentais, dados de CNPJ e informações de parceiros, para criar um data lake abrangente.

Análise de Dados: Criar tabelas OLAP para permitir análises estatísticas e de tendências sobre os dados integrados. Isso envolve a aplicação de algoritmos estatísticos e a identificação de filtros significativos, como quantidade/região, quantidade/vendas por tempo, quantidade de domicílios por região e quantidade de consumo por região/tempo.

Visualização de Dados: Gerar um infográfico que representa visualmente as análises e tendências identificadas de forma clara e concisa. Isso permitirá que os decisores tomem decisões estratégicas informadas com base nas informações apresentadas.

## 2. Objetivos

Nesta seção, apresenta-se os objetivos do projeto que são as metas e resultados esperados a serem alcançados com a execução do mesmo. Servindo como uma referência para orientar as ações do projeto e ajudar a equipe a entender o que precisa ser feito e como avaliar o sucesso do projeto.

### 2.1. Objetivos Gerais

O objetivo deste projeto é criar uma solução eficiente de Big Data para uma empresa distribuidora, utilizando a infraestrutura da AWS. Tendo como foco estabelecer um pipeline de dados escalável para coletar, processar e analisar grandes volumes de informações, transformando-os em insights estratégicos. Além disso, visa-se desenvolver um infográfico que resuma os resultados das análises de forma clara e acessível, melhorando a experiência do usuário na interpretação dos dados e auxiliando na tomada de decisões eficazes. A realização deste projeto irá fornecer à empresa distribuidora uma solução robusta, capaz de extrair valor dos dados e impulsionar sua eficiência operacional, fornecendo uma vantagem competitiva no mercado.

### 2.2. Objetivos Específicos

1. Identificação e integração das fontes de dados relevantes, incluindo dados internos e externos, garantindo a qualidade e confiabilidade das informações coletadas.
2. Desenvolvimento de um pipeline de dados eficiente que permita a ingestão, transformação e limpeza dos dados coletados, garantindo a integridade e consistência das informações ao longo do processo.
3. Implementação de mecanismos de exclusão de dados sensíveis provenientes da API do parceiro, garantindo a conformidade com regulamentações de privacidade e segurança.
4. Análises avançadas dos dados coletados, utilizando técnicas preditivas, prescritivas e diagnósticas para extrair insights estratégicos e identificar padrões relevantes para a empresa distribuidora.

5. Desenvolvimento de um sistema de visualização de dados que traduza os resultados das análises em gráficos e infográficos claros e acessíveis, facilitando a interpretação e compreensão dos insights pelos usuários.

## 2.3. Justificativa

A justificativa para a realização deste projeto reside na importância estratégica de lidar eficientemente com o Big Data. O volume crescente de informações geradas pelas organizações demanda a capacidade de coletar, processar e analisar esses dados para obter insights açãoáveis. A análise de grandes conjuntos de dados é fundamental para a tomada de decisões informadas e para identificar tendências, padrões e oportunidades de negócios. Além disso, a aplicação de tecnologias avançadas, como bancos de dados distribuídos e algoritmos de machine learning, é necessária para enfrentar os desafios atuais de escalabilidade e disponibilidade em ambientes distribuídos. Investir nesse projeto permitirá que a empresa distribuidora aproveite ao máximo seus dados, ganhando uma vantagem competitiva no mercado e impulsionando seu crescimento sustentável.

## 3. Compreensão do Problema

Apresenta-se nessa sessão as descrições das análises voltadas ao desenvolvimento de resultados do projeto, para a empresa Integration, a respeito da construção de um MVP (Produto mínimo viável), exibindo os identificadores de mercado, de acordo com as ferramentas de negócio utilizadas.

### 3.1 Canvas Proposta de Valor

A seguir, apresentamos o Canvas Proposta de Valor, elaborado para a empresa Integration, parceira neste módulo do projeto. Através do Canvas Proposta de Valor, buscamos entender e mapear de maneira estruturada as soluções oferecidas pela Integration, especialmente no que tange ao mercado consumidor de alimentos brasileiro.

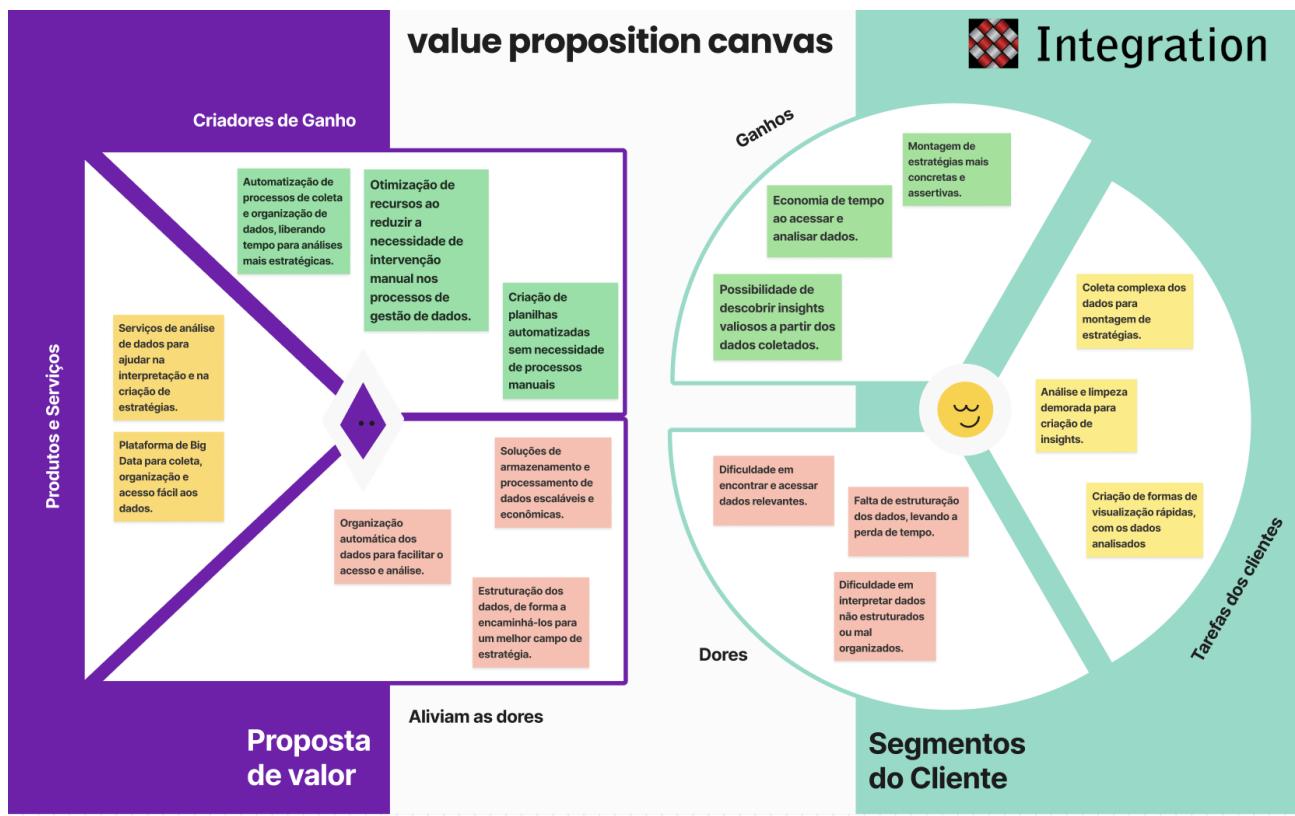


Figura 01: Canvas Proposta de Valor

Fonte: Elaboração própria

## Criadores de Ganho:

Essa seção destaca as principais vantagens e benefícios gerados pela empresa. O projeto foca em automatizar processos de coleta e organização de dados, otimizar recursos ao minimizar intervenções manuais em processos de gestão de dados e criar planilhas automatizadas, reduzindo assim a necessidade de processos manuais.

## Produtos e Serviços:

O Canvas evidencia os principais produtos e serviços oferecidos, como serviços de análise de dados para auxiliar na interpretação e criação de estratégias, plataformas de Big Data para coleta e organização com acesso facilitado, e soluções automatizadas para armazenamento e processamento de dados.

## Aliviam as Dores:

Compreendendo os desafios do mercado, o projeto oferece soluções que amenizam dificuldades enfrentadas pela Integration. Estas incluem organização automática dos dados, estruturação eficiente dos mesmos e sua direção para melhores campos estratégicos, visando superar barreiras como dificuldades de acesso e interpretação de informações.

## **Ganhos:**

A proposta do projeto é entregar valores tangíveis a Integration. Isso é refletido em benefícios como economia de tempo ao acessar e analisar dados, possibilidade de descobrir insights valiosos a partir dos dados coletados e montagem de estratégias mais assertivas.

## **Tarefas do Cliente:**

A integration busca realizar a coleta complexa de dados, fazer análises e limpezas demoradas para criação de insights e desenvolver formas rápidas de visualização.

## **Dores:**

A integration enfrenta desafios como dificuldades em encontrar e acessar dados relevantes, lidar com falta de estruturação dos dados que leva à perda de tempo, e o obstáculo de interpretar dados não estruturados ou mal organizados.

### **3.2. Matriz de Risco**

A matriz de riscos é uma ferramenta que proporciona uma análise ampla das ameaças e oportunidades do projeto. Com ela podemos definir quais são as ameaças com maiores probabilidades e impactos no nosso projeto, além das oportunidades que são vigentes dentro do desenvolvimento. Sua principal função é auxiliar a empresa a tomar decisões baseadas nos impactos e na probabilidade desses riscos acontecerem.

Matriz de Risco											
Probabilidade:	Riscos						Oportunidades				
	Muito Alta	1									
	Alta	2									
	Média	3									
	Baixa	4									
	Muito Baixa	5									
		1	2	3	4	5	5	4	3	2	1
	Muito Baixo	Baixo	Médio	Alto	Muito Alto	Muito Alta	Alta	Média	Baixa	Muito Baixa	
	Impacto										

Figura 02: Matriz de risco

Fonte: Elaboração própria

A Estratégia de Mitigação de Riscos com Descrição e Responsabilidades é um componente fundamental do nosso projeto de Integração, Gerenciamento e Análise de Big Data. Este plano detalha ações específicas para identificar, avaliar e reduzir riscos críticos. Cada risco é analisado com consequências e probabilidades destacadas, enquanto os responsáveis garantem a execução eficaz das ações de mitigação. Este plano fortalece nossa capacidade de tomar decisões informadas e assegura o sucesso do projeto, controlando eventos imprevistos.

Plano de Ação					
Matriz	Descrição do Impacto	Probabilidade	Impacto	Descrição da Ação	Responsável
2 - 3	Mesmo improvável, desafios na integração do pipeline com os sistemas do cliente podem causar atrasos, destacando a importância de uma abordagem proativa.	Alta	Médio	Irá idear uma análise abrangente dos sistemas existentes do cliente, identificando todos os sistemas relevantes e avaliando suas estruturas, funcionamento e dados.	PO e SM da Sprint
2 - 4	Mesmo improvável, possíveis impactos devido a atrasos não úteis poderiam resultar em atrasos e destacam a importância de planejamento flexível.	Alta	Alto	Durante o planning identificar os fatores e pontos facultativos, de modo que a distribuição das tarefas seja realizada levando em consideração o volume de trabalho e a disponibilidade de tempo.	PO e SM da Sprint
2 - 5	Mesmo improvável, a falta de conhecimento na AWS e suas ferramentas tem impacto crítico, enfatizando a importância de medidas de treinamento.	Alta	Muito Alto	Garantir que todos se dedicarem para incluir e preparar sobre a AWS e suas ferramentas. Utilizando os materiais de automação e o auxílio dos professores/orientadores.	SM da Sprint
3 - 3	Em cenários raros, conflitos na implementação podem resultar em atrasos moderados, destacando a necessidade de abordagem proativa.	Média	Médio	Promover discussões no daily para que cada membro possa expressar suas opiniões e argumentos em relação às ferramentas preferidas. Faz uma reunião constativa de ideias, baseada em evidências e experiências anteriores.	Pedro
3 - 4	A falta de engajamento, mesmo improvável, pode levar a atrasos e falta de comprometimento, enfatizando a necessidade de comunicação eficaz.	Média	Alto	Encorajar uma comunicação transparente, incentivando a participação ativa das equipes e oferecer feedback nas reuniões. Além de promover um ambiente de trabalho colaborativo e reconhecer o desempenho contributivo para manter a motivação e o comprometimento dos membros do projeto.	Dayfan
4 - 3	Com probabilidade muito baixa, problemas na qualidade dos dados podem afetar a precisão das análises, enfatizando a necessidade de controle rigoroso de qualidade.	Baixa	Médio	Irá idear a definição de diretrizes detalhadas para a qualidade dos dados de entrada.	Lucas
4 - 4	A ocorrência improvável de atrasos nas entregas das Sprints teria um impacto considerável, enfatizando a necessidade de práticas eficazes de gerenciamento.	Baixa	Alto	Questionar durante a daily o status quo de cada atividade de cada membro do grupo, e indagando quais suas atividades para o dia útil, além disso, antes do final da dev, questionar qual o status das tarefas desempenhadas.	PO e SM da Sprint
4 - 5	A desorganização do GitHub, embora improvável, teria impacto sério, destacando a necessidade de organização rigorosa.	Baixa	Muito Alto	Ser feita uma revisão em grupo de todas as tarefas cumpridas ao decorrer do projeto e, ao final da Sprint, deixar o PO da Sprint responsável por subir os conteúdos da branch de "dev" para "main".	PO da Sprint
5 - 1	Com uma probabilidade muito baixa de ocorrem alterações regulatórias, o impacto seria mínimo, permitindo a continuidade do projeto com estabilidade.	Muito Baixa	Muito Baixo	Monitorar constantemente as mudanças regulatórias relevantes em relação à coleta e uso de dados.	PO e SM da Sprint
5 - 5	Mesmo com baixa probabilidade, o vazamento de dados sensíveis teria um impacto devastador, destacando a importância crítica da proteção contra esse risco.	Muito Baixa	Muito Alto	Desenvolver políticas de privacidade e segurança de dados, garantindo que todos na minha equipe as compreendam e as sigam.	PO da Sprint

Figura 03: Plano de ação

Fonte: Elaboração própria

Para melhor visualização da Matriz de risco → [Link para o FIGMA onde foi desenvolvido.](#)

### 3.3. TAM SAM SOM

O desenvolvimento da análise TAM (Total Addressable Market), SAM (Serviceable Addressable Market) e SOM (Serviceable Obtainable Market) no mercado foi realizado com ênfase na ferramenta em desenvolvimento no âmbito do projeto de Big Data entre a empresa Integration e a faculdade Inteli. Nossa abordagem concentrou-se na análise da receita, um fator crucial para o planejamento estratégico da empresa, e na compreensão do tipo de mercado em que nossa parceira está inserida. Nesse contexto, exploramos conceitos e dados pertinentes ao mercado de Business Intelligence, especificamente no segmento de varejo e serviços alimentares no Brasil. Inicialmente, nosso enfoque abrangeu o mercado nacional como um todo, com uma progressão que nos levou a uma análise da cidade de São Paulo.

TAM	SAM	SOM
<p><b>TAM: Total Addressable Market</b> 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. Aqui estamos representando o mercado de Business Intelligence no Brasil, incluindo as receitas geradas por empresas que atuam nesse setor em todo o país.</p>	<p><b>SAM: Serviceable Available Market</b> <b>Parcela do mercado</b> 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. Corresponde ao segmento de varejo alimentar e de serviço alimentar no Brasil, inserido no TAM. Então, em torno de 40% das empresas desse setor, utilizam de soluções de Business Intelligence.</p>	<p><b>SOM: Serviceable Obtainable Market</b> 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 <b>perspectivas de curto prazo</b> (de 1 a 2 anos). Nessa faixa, analisamos o contexto da indústria alimentícia em São Paulo. Então, pegamos uma faixa do SAM, que era do Brasil todo, e analisamos.</p>

Figura 04: TAM SAM SOM

Fonte: Elaboração própria, em conjunto com outras as equipes da sala

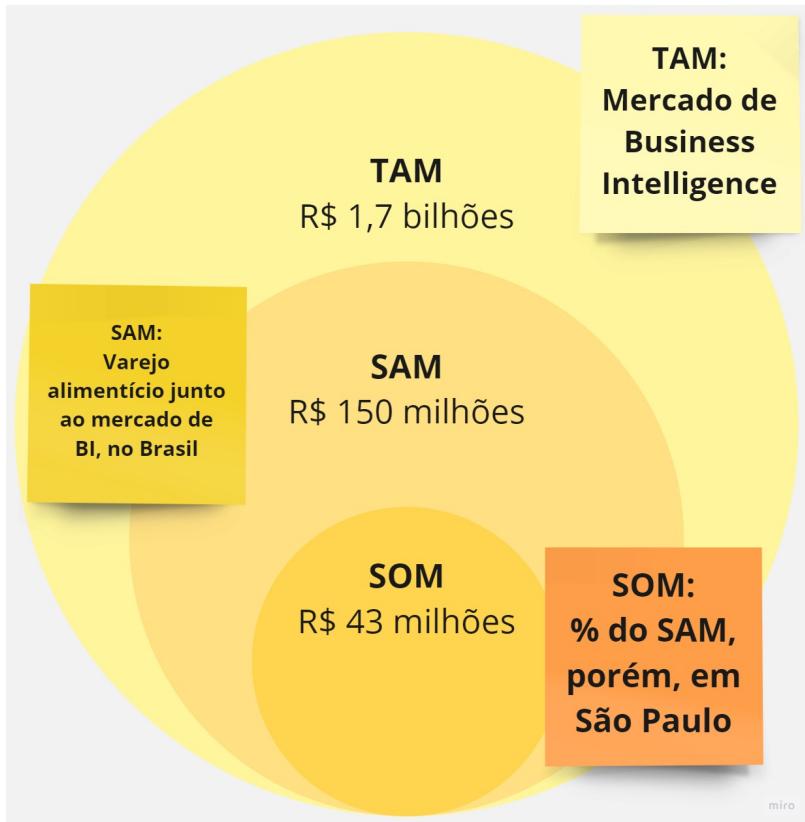


Figura 05: TAM SAM SOM

Fonte: Elaboração própria

	Tamanho estimado do mercado	Notas e comentários
<b>TAM</b>	R\$ 1,7 bilhões	No TAM, estamos avaliando o mercado de Business Intelligence, uma vez que é um setor que a Integration aborda em conjunto com o varejo. A receita gerada está localizada no Brasil.
<b>SAM</b>	R\$ 150 milhões	Nesta análise, estamos conectando o mercado de varejo alimentício com o mercado de BI para identificar a presença conjunta de ambos, ambos situados no Brasil.
<b>SOM</b>	R\$ 43 milhões	Por fim, realizamos uma análise com foco local e direcionada ao nicho do mercado SAM, com maior segmentação, especificamente em São Paulo.

Figura 06: TAM SAM SOM

Fonte: Elaboração própria

## 4. Lean Inception

Nesta seção, apresenta-se o Lean Inception, uma técnica baseada na metodologia ágil que visa definir o escopo e os requisitos do produto de forma colaborativa e eficiente, de todo o time e das partes interessadas na solução.

### 4.1. O Produto (É – Não É – Faz – Não Faz)

Este documento tem como objetivo delimitar as capacidades, limitações e características fundamentais do sistema em desenvolvimento, estabelecendo um melhor entendimento do escopo do projeto, evitando ambiguidades e garantindo uma visão unificada entre os participantes do grupo e com os parceiros do projeto.

Matriz de 4 Quadrantes:

Com essa finalidade em mente, uma matriz com quatro quadrantes é criada, contendo os seguintes conteúdos:

É: Neste quadrante, delineamos tudo aquilo que o produto é, destacando as características essenciais que o definem.

Não é: Este quadrante estabelece limites claros, esclarecendo tudo aquilo que o produto não é. Ao fazê-lo, definimos áreas específicas que não serão abordadas no projeto.

A partir desses dois quadrantes já é possível delimitar um pouco o escopo do sistema, restando os outros dois quadrantes:

Faz: Ao concentrar-nos no que o produto faz, é possível identificar as primeiras funcionalidades da solução, material importante para discutir com o Product Owner do projeto.

Não faz: Identificando o que o produto não faz, comunicamos de forma inequívoca que determinados aspectos não fazem parte do escopo do projeto.

Esta abordagem não apenas economiza tempo ao evitar retrabalhos, mas também contribui para um entendimento compartilhado e alinhado entre todos os envolvidos no desenvolvimento da solução.

#### **4.1.1 O que o Sistema É:**

- Ferramenta de Suporte à Tomada de Decisão: É uma ferramenta projetada para auxiliar na tomada de decisões, oferecendo insights a partir da análise de grandes volumes de dados.
- Personalizável e Configurável: Oferece opções de personalização para atender às necessidades específicas do usuário, permitindo configurações conforme os requisitos da análise.
- Solução para Armazenamento de Dados: É uma solução eficaz para o armazenamento e recuperação eficiente de grandes conjuntos de dados, facilitando análises.
- Compatível com Diversos Bancos de Dados: É compatível com uma variedade de bancos de dados, proporcionando flexibilidade na escolha da infraestrutura de armazenamento a depender do tipo de dado.
- Ambiente Escalável e Adaptável: É projetado para operar em ambientes escaláveis, adaptando-se dinamicamente às necessidades do projeto e ao volume crescente de dados.

#### **4.1.2. O que o Sistema Não É:**

- Plataforma de Desenvolvimento de Modelos de Machine Learning: Não é uma plataforma de desenvolvimento de modelos de machine learning. O foco está na análise e visualização dos dados, não na criação de algoritmos preditivos complexos.
- Sistema de Segurança Independente: Não atua como um sistema de segurança independente. A implementação de medidas de segurança deve ser feita em conjunto com práticas recomendadas de segurança de dados a partir da plataforma cloud em que os dados foram inseridos.
- Ferramenta de Análise de Linguagem Natural (NLP): Não realiza análise avançada de linguagem natural. A interpretação contextual dos dados requer ferramentas especializadas.
- Substituto para Processos de Garantia de Qualidade Manuais: Não substitui processos de garantia de qualidade manuais. A validação e verificação contínua dos dados são responsabilidades essenciais do usuário.

- Solução Única para Todas as Fontes de Dados: Não é uma solução universal para todas as fontes de dados. Requer configurações específicas para integrar eficientemente com diferentes tipos de fontes.

#### **4.1.3. O que o Sistema Faz:**

- Pipeline de Engenharia de Dados Eficiente: Desenvolve um pipeline para a coleta, transformação e carga (ETL) de dados, garantindo a integridade, qualidade e segurança do processo.
- Geração Automática de Infográficos: Automatiza a criação de infográficos a partir dos dados consolidados, facilitando a interpretação visual dos insights pelos usuários.
- Armazenamento em Cubo de Dados: Constrói um cubo de dados para armazenamento, permitindo consultas durante as análises com variáveis pré definidas.
- Adaptação Dinâmica à Escala: Opera em ambientes escaláveis, adaptando-se dinamicamente à demanda de dados diferentes e usuários.
- Integração com Ferramentas de Visualização Externas: Facilita a integração com ferramentas externas de visualização para análises personalizadas.

#### **4.1.4. O que o Sistema Não Faz:**

- Análise Preditiva Avançada: Não realiza análises preditivas avançadas. O foco está na apresentação e interpretação de dados históricos.
- Implementação de Algoritmos de Machine Learning: Não oferece suporte direto à implementação de algoritmos de machine learning. Recomenda-se a integração com plataformas dedicadas nesses casos.
- Verificação Automatizada de Qualidade de Dados: Não inclui uma ferramenta de verificação automatizada de qualidade de dados. A validação precisa ser configurada conforme as necessidades específicas do projeto, sendo esses adicionados pelo usuário desenvolvedor no momento de ingestão dos dados.
- Manipulação Direta de Fontes Externas Complexas: Não manipula diretamente fontes externas complexas sem configuração prévia. A complexidade das fontes pode requerer personalizações específicas.

- Ferramenta de Business Intelligence Completa: Não substitui uma plataforma completa de business intelligence. Enquanto oferece recursos visuais, outras funcionalidades podem necessitar integração com ferramentas especializadas.

## **5. Análise de Experiência do Usuário**

Nesta sessão, apresenta-se a análise de experiência do usuário, a qual através da aplicação de estratégias, visa compreender como os usuários interagem com sistemas, produtos e serviços. O objetivo é melhorar a satisfação e a eficiência dessas interações, levando em conta aspectos subjetivos como emoções, percepções e expectativas dos usuários.

### **5.1. Personas**

As personas são representações detalhadas, porém fictícias, dos clientes ideais que a solução busca atender, além de ocupar um papel crucial ao compreender e orientar projetos e soluções. Tendo em vista nosso atual projeto, que envolve o desenvolvimento de uma consultoria de marketing e vendas baseada em Big Data, criamos dois diferentes tipos de personas para facilitar a compreensão, de forma visual, dos usuários que utilizarão nossa solução: a gerente de Marketing e Vendas e o Analista de Dados.

Essas personas são criadas com base nos setores-chave que garantem a eficácia da solução. Onde, cada uma delas contém características, comportamentos e preferências alinhados com o contexto da Integration. Além de ajudar a equipe a visualizar os usuários finais, essas personas também são utilizadas para definir estratégias e recursos que atendam às carências e expectativas de cada perfil.

## Ana Mercadante



Ana tem uma rotina agitada, equilibrando o trabalho com sua família. Comprometida com o sucesso de sua equipe, busca maneiras de melhorar o desempenho das vendas e o impacto das estratégias de marketing. No entanto, ela vê horas perdidas na criação manual de planilhas para acessar dados de consumo, regiões e possíveis canais de venda das planilhas.

**"Simplificar a análise de dados economizaria meu tempo e minha sanidade."**

**Idade:** 35 anos  
**Profissão:** Gerente  
**Setor:** Marketing e Vendas  
**Família:** Casada com dois filhos  
**Localização:** São Paulo, SP  
**Perfil socioeconômico:** Classe média

**Personalidade**

- Extrovertida
- Analítica
- Proativa

**Nível de Letramento Digital**

Sabe utilizar aplicações web e arquivos csv, mas não sabe utilizar soluções com programação, mesmo de alto nível.

**Dores, necessidades e desejos**

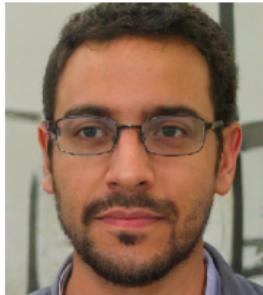
- Criação manual de planilhas.
- Suscetibilidade a erros na abordagem manual.
- Frustração de lidar com a duplicação de dados.
- Dificuldade em consultar dados como regiões em potencial, consumo, etc.
- Necessidade de simplificar o processo de cruzamento de dados.
- Busca por soluções a fim de facilitar a tomada de decisões estratégicas.
- Otimizar o desempenho do setor de marketing e vendas.

**Motivação**  
Adquirir tecnologias que auxiliem tomada de decisão

Figura 07: Persona - Gerente Marketing e Vendas.

Fonte: Elaboração própria.

## Vinicius Monteiro



Vinicius é um indivíduo tranquilo e reservado, que prefere trabalhar de forma independente e concentrada. Ele é altamente analítico e adora resolver problemas complexos. Sua natureza meticolosa o torna um perfeccionista, entregando resultados precisos. Portanto, se encontra irritado por conta do trabalho repetitivo da criação de planilhas manualmente. Sendo assim, uma automatização do pipeline de Big Data eficiente resloveria o seu problema atual.

**"Automatizar o pipeline de Big Data deixaria meu trabalho mais fácil e eficiente."**

**Idade:** 28 anos  
**Profissão:** Técnico de TI  
**Setor:** Dados  
**Família:** Casada sem filhos  
**Localização:** São Paulo, SP  
**Perfil socioeconômico:** Classe média

**Personalidade**

- Introvertido
- Perfeccionista
- Workaholic

**Nível de Letramento Digital**

Sólido conhecimento no uso de serviços e recursos AWS. Habilidoso em configurar e gerenciar pipelines de Big Data,

**Dores, necessidades e desejos**

- Necessidade de um pipeline de Big Data eficiente.
- Obtenção de análises estatísticas avançadas.
- Infográfico intuitivo e atraente.
- Otimização de desempenho do pipeline.
- Automatização de planilhas, antes realizadas manualmente.

**Motivação**  
Inovação tecnológica e entrega de valor

Figura 08: Persona - Analista de Dados

Fonte: Elaboração própria

## 5.2. Jornada do Usuário

Para melhor ilustrar as etapas do relacionamento do usuário com um produto ou uma solução, abordando a experiência antes, durante e após o seu uso, em relação a um cenário específico. A dinâmica do usuário com a solução é importante para registrar a interação do usuário ao longo do tempo, identificando pontos integrados na narrativa, permitindo a compreensão dessa interação, destacando os pontos de atenção.

Duas jornadas foram elaboradas, para cada persona, a fim de contextualizar os eventos que ocorrem desde o recebimento da solicitação, descrita no cenário, até o momento em que desempenham suas funções.

Jornada de Ana Mercadante		Expectativas		
FASE 1: Conscientização	FASE 2: Consideração	FASE 3: Uso	FASE 4: Contratempo	FASE 5: Resultado
Ana recebeu o projeto, e reconhece a necessidade conseguir acessar todos os dados relevantes sobre o mercado do cliente, o setor em que atua, o censo demográfico.  <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <i>"Preciso acessar os dados para diagnosticar o problema do cliente da melhor maneira possível."</i> </div> Sentimento: Neutro	Após analisar a natureza dados específicos do cliente, Ana decide adotar a solução hospedada em AWS, que agrupa informações do cliente e outros dados públicos, criando um ambiente propício para a análise desses dados em uma base segura.  <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <i>"Esta solução parece a mais adequada para atender às nossas necessidades de análise de dados."</i> </div> Sentimento: Positivo	O uso da ferramenta possibilita a análise do setor do cliente, potencial de consumo, possíveis canais de venda e potenciais regiões, definindo um plano, mapeando públicos e o investimento necessário para assegurar o sucesso da estratégia.  <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <i>"Assim posso personalizar estratégia, assegurando a disponibilidade de produtos aos consumidores."</i> </div> Sentimento: Positivo	Devido ao tempo considerável que Ana precisa investir na geração dos infográficos, essa limitação impacta significativamente o uso da solução para realizar simulações de diferentes cenários. Além disso, essa demora na produção dos infográficos pode interromper o fluxo de pensamento de Ana e prejudicar a continuidade de sua análise e tomada de decisões.  <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <i>"Toda vez que usamos essa solução, demora um tempão... Deveria ser mais eficiente."</i> </div> Sentimento: Negativo	Ana obtém os resultados, desenvolvendo uma estratégia comercial direcionada com base em insights sobre os dados. Apesar dos contratemplos, Ana supera os desafios, adaptando sua estratégia e finalizando o projeto e encerrando a jornada.  <div style="border: 1px solid #ccc; padding: 5px; width: fit-content;"> <i>"Uhuuu! O planejamento foi realizado, faltando só implementar."</i> </div> Sentimento: Positivo

### Oportunidades

Dentre as oportunidades, é possível destacar a melhoria na apresentação dos dados, garantindo que Ana tenha acesso claro às funcionalidades e benefícios da solução. Aprimoramentos na interface da solução podem permitir análises mais detalhadas, resultando em um investimento mais preciso de recursos, com alto desempenho da equipe ao atuar nos projetos.

### Picos e Vales

Quando Ana acessa a solução, por cruzar dados, proporciona uma visualização por infográficos, dando suporte nas análises e a possibilidade de usar as informações geradas em simulações, representando o maior pico da jornada especificada. Por outro lado, pelo tamanho da base de dados gerar informações leva tempo, gerando uma limitação no seu uso, sendo esse o maior vale da jornada especificada.

### 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. Importante também realizar simulações da estratégia planejada, utilizando a solução de análise de dados como uma ferramenta contínua de otimização.

### Touchpoints

1. Aplicação hospedada na AWS: Permite acesso a solução em computadores via aplicação web.
2. API: A solução pode ser acessada por meio de uma API, permitindo integrações personalizadas e o uso em outras ferramentas.
3. E-mail: Ana pode receber relatórios ou notificações por e-mail da solução, mantendo-a informada sobre atualizações.

miro

Figura 09: Jornada do usuário de Ana Mercadante. Fonte: Autoria própria.



### Jornada de Vinicius Monteiro

Vinicius precisa de informações específicas no cubo de dados para a equipe de marketing e vendas. A falta dos filtros obstrui análises sobre regiões potenciais de interesse, e o cruzamento das informações do censo, com o consumo das pessoas dessas regiões.

### Expectativas

É esperado que a implementação de filtros nos serviços da AWS permita o acesso a informações específicas do cubo de dados, de forma que a equipe de marketing e vendas prossiga no desenvolvimento do projeto para o cliente.

FASE 1: Conscientização	FASE 2: Consideração	FASE 3: Uso	FASE 4: Contratempo	FASE 5: Resultado
Vinicius toma consciência da necessidade de adicionar as informações especificadas no cubo de dados. Ele percebe que a falta de filtros está dificultando a análise de regiões de interesse e o cruzamento de dados do censo.  "A equipe de marketing e vendas precisa dessas informações para identificar os 'regiões' de interesse."	Durante a fase de consideração, Vinicius explora as decisões relacionadas às configurações dos serviços da AWS que serão necessárias para implementar os filtros no cubo de dados.  "Encontrei as configurações para os filtros. Preciso ter certeza de que essa é a escolha certa"	Vinicius começa a ajustar o sistema, e consegue adicionar os dados das regiões de interesse e cruzar informações do censo, juntamente com os dados de consumo.  "Finalmente estou conseguindo isolar os dados das regiões de interesse. Isso é incrível!"	Porém, devido a mudanças nas políticas de segurança, algumas das configurações dos serviços da AWS estão causando conflitos com as novas políticas de permissões. Isso gera atrasos e frustração, uma vez que ele precisa revisar e adaptar as configurações para garantir a conformidade com as políticas atualizadas.  "Ah não! Vou precisar rever todos os conflitos antes de avançar..."	Vinicius consegue adicionar as informações necessárias para a equipe de marketing e vendas. E com a adição dos filtros a jornada do usuário se encerra.  "Sucesso! A equipe de marketing e vendas tem as informações de que precisa."

### Oportunidades

Como a solução está hospedada na nuvem da AWS, é vital garantir a conformidade com as políticas de uso dessa plataforma, garantindo a integridade dos dados e a segurança das informações. Outro aspecto é a modularidade da solução, capaz de cruzar informações indicadas pelo setor de marketing e vendas, e implementar filtros com diferentes vertentes de dados.

### Picos e Vales

O melhor momento dessa jornada é a unificação dos dados, de forma automatizada, possibilitando ajustes nas configurações e variáveis, já que os dados já foram tratados previamente, sendo esse o pico da experiência.

Por outro lado, alterações nas políticas de armazenamento, gerenciamento e segurança dos dados pode levar a uma demora para a resolução de problemas, se tratando de uma outra empresa que armazena esses dados, sendo esse o maior vale da jornada.

### Responsabilidades

É necessário assegurar que as políticas internas e regulamentações externas sejam atendidas, bem como garantir a segurança e conformidade dos dados. Ele deve documentar as configurações e políticas implementadas, solucionar contratemplos e colaborar com equipes relacionadas além da implementação de medidas de segurança, treinamento da equipe e a busca contínua por melhorias na solução.

### Touchpoints

- AWS Management Console para Serviços Gerenciados:** Para serviços gerenciados específicos, a AWS oferece consoles dedicados que fornecem interfaces personalizadas para gerenciamento simplificado.
- Documentação Online:** A documentação oficial da AWS é uma ótima fonte de informações para orientações sobre configurações e boas práticas.
- Suporte Técnico da AWS:** Para problemas mais sérios, é possível entrar em contato com o suporte técnico da AWS para obter assistência.

miro

Figura 10: Jornada do usuário de Vinicius Monteiro. Fonte: Autoria própria.

## 5.3. User Stories

Pode-se definir *User Stories* como descrições simplificadas das funcionalidades possíveis que o usuário possui e deseja dentro da aplicação, escrita com a visão dele. Além de transparecer como o sistema espera alcançar tais objetivos. As tabelas abaixo estão divididas em 6 partes: Número, Título, Personas, História, Critérios de Aceitação e Testes de Aceitação. O número e título servem para identificação, já as personas servem para associar a quem a história pertence. Os dois últimos tópicos descrevem quais são os critérios que aquele usuário deve passar no sistema para realizar a ação descrita na "história", já o teste diz como o sistema deve agir de acordo com o critério estipulado.

## User Story 1: Configuração de Ambiente AWS pela Equipe de TI

Persona: Equipe de TI

História: Como membro da equipe de TI, desejo configurar o ambiente AWS e seus componentes, a fim de estabelecer a infraestrutura para armazenar, processar e manter os dados do projeto de maneira segura e eficiente.

Critério de avaliação: Configuração da Infraestrutura AWS

- 1.1 Condição de Aceite: A infraestrutura da AWS está configurada e pronta para uso.
- 1.2 Condição de Recusa: A infraestrutura da AWS não está configurada ou não está pronta para uso.

*Teste de Aceitação - Critério 1: Configuração de Ambiente AWS*

- 1.1 {Aceito}
  - Explicação do Aceito: A infraestrutura AWS foi configurada com todos os componentes funcionando conforme esperado, e a equipe de TI pode acessar e utilizar o ambiente para processar e armazenar dados.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na configuração da infraestrutura AWS, como falhas de implementação, configurações incorretas ou incapacidade de acessar o ambiente AWS.

*Teste de Aceitação - Critério 2: Verificação de Acesso ao Ambiente AWS*

- 2.1 {Aceito}
  - Explicação do Aceito: A equipe de TI pode acessar o ambiente AWS com as credenciais fornecidas e realizar ações básicas, como listar serviços disponíveis.
- 2.2 {Recusado}
  - Explicação do Recusado: A equipe de TI não consegue acessar o ambiente AWS com as credenciais fornecidas ou enfrenta dificuldades técnicas que impedem o acesso

### *Teste de Aceitação - Critério 3: Funcionamento dos Componentes Essenciais*

- 3.1 {Aceito}
  - Explicação do Aceito: Todos os componentes essenciais, como servidores, bancos de dados e serviços de armazenamento, estão funcionando conforme esperado.
- 3.2 {Recusado}
  - Explicação do Recusado: Alguns dos componentes essenciais estão com problemas ou não estão funcionando conforme esperado

### *Teste de Aceitação - Critério 4: Teste de Resiliência e Redundância*

- 4.1 {Aceito}
  - Explicação do Aceito: A infraestrutura AWS passou em testes de resiliência e redundância, garantindo a disponibilidade contínua dos serviços.
- 4.2 {Recusado}
  - Explicação do Recusado: A infraestrutura AWS não passou nos testes de resiliência e redundância, demonstrando vulnerabilidades na disponibilidade dos serviços.

Notas : Esta user story é o primeiro passo para as demais user stories, especialmente a User Story 2 (Ingestão de Dados no Datalake com Ênfase na Segurança).

Prioridade	Estimativa	Relação
Alta	4 dias	N/A

## **User Story 2: Ingestão de Dados no Datalake com Ênfase na Segurança**

Persona: Analista de Dados

História: Como equipe de Dados, desejo realizar a ingestão de dados no datalake da AWS S3, com foco na segurança e controle de acesso, para garantir a integridade dos dados, a confidencialidade das informações e a segregação dos dados sensíveis.

## Critério de avaliação: Ingestão de dados no Datalake

- 1.1 Condição de Aceite: Os dados foram corretamente injetados no datalake AWS S3 com medidas de segurança adequadas, incluindo confidencialidade, integridade e controle de acesso.
- 1.2 Condição de Recusa: A ingestão de dados não foi realizada de forma segura, ou as medidas de segurança, confidencialidade e controle de acesso não foram aplicadas conforme necessário, ou houver risco de exposição de dados sensíveis.

### *Teste de Aceitação - Critério 1: Ingestão de Dados no Datalake*

- 1.1 {Aceito}
  - Explicação do Aceito: Os dados foram injetados mantendo a integridade dos mesmos, com medidas de segurança aplicadas, e o acesso é controlado, no datalake AWS S3.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na ingestão de dados, como dados incorretos, perda de integridade ou falta de medidas de segurança adequadas, confidencialidade ou controle de acesso.

### *Teste de Aceitação - Critério 2: Confidencialidade de Dados*

- 1.1 {Aceito}
  - Explicação do Aceito: Os dados foram injetados com confidencialidade no datalake AWS S3, garantindo que informações sensíveis não sejam acessíveis sem autorização adequada.
- 1.2 {Recusado}
  - Explicação do Recusado: Os dados foram injetados sem confidencialidade, com risco de exposição de informações sensíveis.

### *Teste de Aceitação - Critério 3: Controle de Acesso ao Docker e VPC*

- 1.1 {Aceito}
  - Explicação do Aceito: O Docker e a infraestrutura de acesso ao datalake foram configurados dentro de uma VPC (Virtual Private Cloud) com acesso controlado, garantindo que somente usuários autorizados tenham acesso aos dados.
- 1.2 {Recusado}

- Explicação do Recusado: O Docker e a infraestrutura não foram configurados dentro de uma VPC, colocando em risco a segurança e a confidencialidade dos dados, permitindo acesso não autorizado.

Notas: Esta história se concentra na ingestão de dados iniciais para testes.

Prioridade	Estimativa	Relação
Alta	14 dias	User Story 1

### User story 3 - Análise Inicial dos Dados

Persona: Analista de Dados

História: Como analista de dados desejo realizar uma análise inicial dos dados no datalake recém-ingestado, a fim de compreender a qualidade e relevância dos dados para as futuras análises.

Critério de avaliação: Análise Inicial dos Dados

- 1.1 Condição de Aceite: O analista de dados concluiu a análise inicial, documentando a qualidade dos dados com métricas de integridade, precisão, consistência, e sua relevância para as futuras análises.
- 1.2 Condição de Recusa: Problemas críticos na análise dos dados, falta de documentação da qualidade dos dados ou falta de relevância dos dados para análises futuras.

*Teste de Aceitação - Critério 1: Análise de Qualidade dos Dados*

- 1.1 {Aceito}
  - Explicação do Aceito: A análise inicial dos dados foi documentada com métricas de qualidade dos dados, incluindo integridade, precisão e consistência.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na análise de qualidade dos dados, como métricas inaceitáveis de integridade, precisão ou consistência.

- *Teste de Aceitação - Critério 2:Relevância dos Dados para Análises Futuras*
- 1.1 {Aceito}
  - Explicação do Aceito: Os dados analisados foram considerados relevantes para análises futuras, com base na documentação da equipe de Análise de Dados.
- 1.2 {Recusado}
  - Explicação do Recusado: Os dados analisados não são relevantes para análises futuras, conforme documentado pela equipe de Análise de Dados.
- *Teste de Aceitação - Critério 3: Documentação da Análise*
- 1.1 {Aceito}
  - Explicação do Aceito: A análise inicial dos dados foi devidamente documentada, incluindo as métricas de qualidade dos dados, sua quantidade e a avaliação de relevância para análises futuras do projeto.
- 1.2 {Recusado}
  - Explicação do Recusado: Falta de documentação adequada da análise inicial dos dados, dificultando a avaliação de qualidade e relevância.

Notas : Esta história se concentra na análise estatística inicial dos dados para identificação de tendências.

Prioridade	Estimativa	Relação
Média	7 dias	User Story 1 - User Story 2 -

## **User Story 4 - Análise e Visualização Preliminar de Dados Governamentais e de CNPJ no Datalake para Identificar Necessidades de Tratamento ETL**

Persona: Analista de Dados

História: Como analista de dados, desejo implementar uma visualização que permita a análise preliminar de todos os dados já recebidos no data lake, incluindo dados governamentais e de CNPJ, a fim de identificar a necessidade de tratamento ETL.

Critério de avaliação: Análise e Visualização Preliminar de Dados

- 1.1 Condição de Aceite: A visualização preliminar foi implementada no datalake, permitindo a análise de todos os dados já recebidos, identificando as necessidades de tratamento ETL.
- 1.2 Condição de Recusa: Problemas críticos na implementação da visualização preliminar que não permitem a análise dos dados ou a identificação das necessidades de tratamento ETL.

*Teste de Aceitação - Critério 1: Implementação da Visualização Preliminar de Dados*

- 1.1 {Aceito}
  - Explicação do Aceito: A visualização preliminar foi implementada conforme especificado, permitindo a análise de todos os dados já recebidos no datalake.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na implementação da visualização preliminar impedem a análise dos dados.
- *Teste de Aceitação - Critério 2: Identificação de Necessidades de Tratamento ETL*
- 1.1 {Aceito}
  - Explicação do Aceito: A análise preliminar dos dados identificou as necessidades de tratamento ETL, conforme necessário.
- 1.2 {Recusado}
  - Explicação do Recusado: A análise preliminar não foi capaz de identificar adequadamente as necessidades de tratamento ETL devido a problemas na implementação.
- *Teste de Aceitação - Critério 3: Insights Iniciais para Relações de Dados*
- 1.1 {Aceito}
  - Explicação do Aceito: A análise preliminar dos dados forneceu insights iniciais sobre como os dados podem ser relacionados para calcular as métricas desejadas, como quantidade de vendas por tempo, quantidade de domicílio por região e quantidade de consumo por região/tempo.
- 1.2 {Recusado}

- Explicação do Recusado: Falta de insights iniciais sobre a relação entre os dados devido a problemas na implementação da visualização preliminar.

Notas: Queremos obter insights iniciais para compreender como esses dados podem ser relacionados para calcular métricas como quantidade de vendas por tempo, quantidade de domicílio por região e quantidade de consumo por região/tempo, conforme requisitado pelo cliente.

Prioridade	Estimativa	Relação
Média	9 dias	User Story 1 - User Story 2 - User Story 3

## **User Story 5 - Importação de Dados Governamentais no Datalake via Pacote Python**

Persona: Analista de Dados

História: Como analista de dados, desejo implementar um processo de importação de dados governamentais no datalake usando um pacote Python para enriquecer nossos dados com fontes governamentais.

Critério de avaliação:

- 1.1 Condição de Aceite: O processo de importação de dados governamentais foi implementado e os dados governamentais estão disponíveis no datalake.
- 1.2 Condição de Recusa: Problemas críticos na implementação do processo ou indisponibilidade dos dados governamentais no datalake.

*Teste de Aceitação - Critério 1: Implementação do Processo de Importação de Dados Governamentais*

- 1.1 {Aceito}

- Explicação do Aceito: O processo de importação de dados governamentais foi implementado e os dados estão disponíveis no datalake.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na implementação do processo e/ou falta de disponibilidade dos dados governamentais no datalake.
- *Teste de Aceitação - Critério 2: Qualidade dos Dados Importados*
- 1.1 {Aceito}
  - Explicação do Aceito: Os dados governamentais importados mantêm a qualidade e integridade necessárias para uso nas análises de dados.
- 1.2 {Recusado}
  - Explicação do Recusado: Os dados importados apresentam problemas críticos de qualidade ou integridade, tornando-os inadequados para análise.

Notas: O critério de tempo ainda será adicionado com o que deve ser esperado como resposta, visto que o time ainda não sabe qual o tempo adequado de resposta do processamento dos dados dentro de um Datalake AWS S3.

Prioridade	Estimativa	Relação
Alta	14 dias	User Story 1 - User Story 2 - User Story 3

## User Story 6 - Implementação de Filtros para o Cubo de Dados

Persona: Analista de Dados

História: Como analista de dados, desejo implementar filtros para o cubo de dados, proporcionando aos usuários a capacidade de refinar e personalizar a visualização de dados de acordo com suas necessidades.

Critério de avaliação:

- 1.1 Condição de Aceite: Os filtros foram implementados corretamente, permitindo que os usuários refinem a visualização de dados de acordo com suas necessidades de forma prática e eficiente.

- 1.2 Condição de Recusa: Problemas críticos na implementação dos filtros e/ou falta de funcionalidade esperada

#### *Teste de Aceitação - Critério 1: Implementação dos Filtros no Cubo de Dados*

- 1.1 {Aceito}
  - Explicação do Aceito: Os filtros foram implementados corretamente e oferecem uma experiência positiva aos usuários. Os usuários podem refinar a visualização de dados com facilidade e eficiência, atendendo às suas necessidades.
- 1.1.2 {Aceitação Adicional}
  - Critério de Aceitação Adicional: O analista de dados consegue aplicar filtros de forma intuitiva utilizando SQL e retornando os resultados da filtragem precisos e a performance da aplicação não é prejudicada.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos na implementação dos filtros e/ou falta de funcionalidade esperada. Os filtros não funcionam conforme o esperado, o analista tem dificuldade em usá-los ou a performance da aplicação é prejudicada.

Notas: A implementação de filtros visa melhorar a usabilidade e a capacidade de personalização da visualização de dados no cubo. Os filtros devem ser intuitivos, eficazes e não devem prejudicar a performance da aplicação. O sucesso da implementação é determinado pela capacidade dos usuários de refinar a visualização de dados de acordo com suas necessidades de forma prática e eficiente.

Prioridade	Estimativa	Relação
Alta	7 dias	User Story 1 - User Story 4 - User Story 5

### **User story 7 - Plataforma de Visualização com Infográfico**

Persona: Analista de Dados

História: Como analista de dados, desejo criar uma plataforma de visualização que inclui um infográfico, desenvolvida no Grafana, para apresentar os dados coletados no cubo de

dados de forma acessível e informativa para o Analista de Marketing e Vendas, que pode não possuir letramento digital.

#### Critério de avaliação: Plataforma de Visualização Criada

- Condição de Aceite: A plataforma de visualização foi desenvolvida com êxito e inclui o infográfico com os dados apropriados.
- 1.2 Condição de Recusa: Problemas críticos no desenvolvimento da plataforma e/ou falta do infográfico esperado.

#### *Teste de Aceitação - Critério 1: Desenvolvimento da Plataforma de Visualização*

- 1.1 {Aceito}
  - Explicação do Aceito: A plataforma de visualização foi desenvolvida conforme especificado e inclui o infográfico com dados relevantes. O Analista de Marketing e Vendas é capaz de acessar e compreender facilmente as informações apresentadas, mesmo que não tenha letramento digital.
- 1.1.2 {Aceitação Adicional}
  - Explicação do Aceito Adicional: A plataforma deve ser acessível por um público não tecnicamente qualificado, e o infográfico deve comunicar efetivamente as informações do cubo de dados de maneira informativa.
- 1.2 {Recusado}
  - Explicação do Recusado: Problemas críticos no desenvolvimento da plataforma ou falta do infográfico esperado. A plataforma não foi desenvolvida conforme as especificações ou o infográfico não atende às necessidades do Analista de Marketing e Vendas.
- *Teste de Aceitação - Critério 2: Acessibilidade para Usuários Não Tecnicamente Qualificados*
- 1.1 {Aceito}
  - Explicação do Aceito: A plataforma é acessível e comprehensível para usuários não tecnicamente qualificados, como o Analista de Marketing e Vendas. A interface e o infográfico são intuitivos e fáceis de usar.
- 1.2 {Recusado}
  - Explicação do Recusado: A plataforma não é acessível ou comprehensível para usuários não tecnicamente qualificados. O Analista de Marketing e

Vendas enfrenta dificuldades ao utilizar a plataforma ou ao interpretar o infográfico.

- *Teste de Aceitação - Critério 3: Compreensão dos Dados pelo Infográfico*
- 1.1 {Aceito}
  - Explicação do Aceito: O infográfico comunica as informações do cubo de dados de forma informativa. O Analista de Marketing e Vendas consegue entender facilmente os dados apresentados.
- 1.2 {Recusado}
  - Explicação do Recusado: O infográfico não comunica as informações do cubo de dados e/ou não permite que o Analista de Marketing e Vendas compreenda os dados apresentados.

Prioridade	Estimativa	Relação
Alta	N+T (Quantidade de trabalho + Tempo) - Necessário ver com a Integration	User Story 6

## **User story 8 - Acesso Restrito para o Analista de Marketing e Vendas**

Persona: Analista de Marketing e Vendas

História: Como Analista de Marketing e Vendas, desejo que meu acesso à plataforma de visualização seja restrito, permitindo-me apenas visualizar o infográfico. Isso garantirá que não possa criar ou acessar o cubo de dados ou análises estatísticas feitas no Ensemble.

Critério de avaliação: Acesso Restrito à Plataforma de Visualização

- 1.1 Condição de Aceite: O Analista de Marketing e Vendas não tem permissão para criar ou acessar o cubo de dados ou análises estatísticas avançadas.
- 1.2 Condição de Recusa: O Analista de Marketing e Vendas tem acesso irrestrito à plataforma.

*Teste de Aceitação - Critério 1: Verificação de Acesso Restrito*

- 1.1 {Aceito}
  - Explicação do Aceito: O Analista de Marketing e Vendas não tem permissão para criar e/ou acessar o cubo de dados e/ou análises estatísticas avançadas, tendo acesso somente à visualização do infográfico.
- 1.2 {Recusado}
  - Explicação do Recusado: O Analista de Marketing e Vendas possui acesso irrestrito, permitindo a criação e/ou acesso ao cubo de dados e/ou análises estatísticas avançadas.

Prioridade	Estimativa	Relação
Alta	7 dias	N/A

## User story 9 - Visualização de Variáveis para Análise do Analista de Marketing e Vendas

Persona: Analista de Marketing e Vendas

História: Como Analista de Marketing e Vendas, desejo ter a capacidade de visualizar todas as variáveis disponíveis no cubo de dados, incluindo quantidade de vendas por tempo, quantidade de domicílio por região e quantidade de consumo por região/tempo, para que eu possa cruzá-las de acordo com as necessidades da análise.

Critério de avaliação: Visualização de Variáveis que estão contidas no Cubo de Dados

- 1.1 Condição de Aceite: O Analista de Marketing e Vendas consegue visualizar todas as variáveis disponíveis no cubo de dados, incluindo quantidade de vendas por tempo, quantidade de domicílio por região e quantidade de consumo por região/tempo.
  - 1.2 Condição de Recusa: O Analista de Marketing e Vendas não consegue visualizar todas as variáveis e/ou enfrenta problemas críticos na visualização.
- Teste de Aceitação - Critério 1: Verificação de Visualização de Variáveis no Cubo de Dados*
- 1.1 {Aceito}

- Explicação do Aceito: O Analista de Marketing e Vendas consegue ver as variáveis disponíveis para cruzamento de forma clara e acessível na plataforma de visualização do Grafana.
- 1.2 {Recusado}
  - Explicação do Recusado: O Analista de Marketing e Vendas não consegue identificar as variáveis disponíveis para cruzamento e/ou a apresentação das variáveis não está clara na plataforma.

Prioridade	Estimativa	Relação
Alta	7 dias	User Story 8

## User story 10 - Tempo de Resposta Rápido para a Aplicação

Persona: Analista de Marketing e Vendas

História: Como Analista de Marketing e Vendas, desejo que a aplicação tenha um tempo de resposta rápido para que eu possa realizar análises de dados de forma eficiente.

Critério de avaliação:

- 1.1 Condição de Aceite: A aplicação responde de forma quase instantânea às solicitações do Analista de Marketing e Vendas, garantindo que análises de dados possam ser realizadas de maneira eficiente e sem atrasos perceptíveis.
- 1.2 Condição de Recusa: A aplicação apresenta atrasos perceptíveis e não atende ao requisito de resposta quase instantânea, prejudicando a eficiência do Analista de Marketing e Vendas em suas análises de dados.

*Teste de Aceitação - Critério 1: Verificação do Tempo de Resposta da Aplicação*

- 1.1 {Aceito}
  - Explicação do Aceito: O Analista de Marketing e Vendas realiza uma solicitação na aplicação para acessar e analisar dados. A aplicação responde de forma quase instantânea e o analista consegue realizar análises de dados de maneira eficiente e sem frustrações devido a atrasos na resposta.
- 1.2 {Recusado}

- Explicação do Recusado: O Analista de Marketing e Vendas realiza uma solicitação na aplicação para acessar e analisar dados. No entanto, a aplicação não responde de forma quase instantânea e apresenta atrasos perceptíveis.

Notas: Para garantir o tempo de resposta rápido, será implementada uma tabela OLAP entre o cubo de dados e a análise estatística Ensemble. Essa abordagem acelera o consumo de dados, permitindo que a aplicação seja ágil e atenda às expectativas do Analista de Marketing e Vendas.

<b>Prioridade</b>	<b>Estimativa</b>	<b>Relação</b>
média	3 dias	User Story 8 - User Story 9

## **6. Análise das Fontes de Dados Disponibilizadas**

A análise das fontes de dados é importante para compreender a origem, o formato, o tamanho e a frequência de atualização dos dados, garantindo a integridade e confiabilidade das informações utilizadas. Essa prática assegura também a um controle melhor dos resultados obtidos, evitando discrepâncias nos dados usados na ingestão, possibilitando também a otimização de processos, identificar tendências e antecipar problemas.

### **6.1. Instituto Brasileiro de Geografia e Estatística (IBGE)**

O Instituto Brasileiro de Geografia e Estatística (IBGE) disponibiliza seus dados por meio do Plano de Dados Abertos (PDA), conforme estabelecido pelo Decreto nº 8.777/2016. O PDA abrange os períodos de 2016-2017, 2018-2019 e 2020-2022. Este documento orienta a implementação e manutenção de processos institucionais para a divulgação de dados abertos, promovendo transparência. A fonte de dados abrange informações estatísticas e geoespaciais, alinhando-se a normativas como a Lei de Acesso à Informação (LAI) e a Infraestrutura Nacional de Dados Abertos (INDA). O IBGE também segue compromissos da Parceria para Governo Aberto. O PDA é um instrumento de planejamento interno, disponibilizado ao público no canal de transparência ativa, e é atualizado a cada biênio, podendo ser adaptado conforme novas diretrizes institucionais e legislações vigentes.

Os dados são disponibilizados por meio do Plano de Dados Abertos (PDA) onde os conjuntos de dados são apresentados em formatos abertos, como CSV e JSON, e são catalogados no Portal Brasileiro de Dados Abertos, seguindo os Padrões de Interoperabilidade de Governo Eletrônico (ePING) e a Cartilha Técnica para Publicação de Dados Abertos no Brasil. A frequência de atualização dos dados varia de acordo com o período estabelecido no PDA, sendo este atualizado a cada biênio, possibilitando a adaptação às novas diretrizes institucionais e legislações vigentes.

### **6.2. Pesquisa de orçamento familiar (POF)**

Pesquisa de Orçamentos Familiares (POF): A fonte de dados da Pesquisa de Orçamentos Familiares (POF) apresenta tabelas referentes à evolução dos Indicadores não Monetários de Pobreza e Qualidade de Vida no Brasil. Estas tabelas, disponíveis em formatos XLS e ODS, abrangem os períodos de 2008-2009 e 2017-2018. Importante notar

que essas estatísticas são consideradas experimentais, estando em fase de teste e avaliação, exigindo cautela em sua utilização. Desenvolvidas para envolver usuários e partes interessadas, as tabelas são atualizadas a cada 5 anos ou mais.

### **6.3. RAIS e CAGED Microdados**

Os Microdados provenientes das bases RAIS e CAGED oferecem informações não identificadas do CAGED Estatístico, onde todos são disponibilizados em formato texto (.txt). Esses conjuntos de microdados, concebidos para integração em aplicativos estatísticos como SPSS, SAS ou R, são regularmente atualizados, sendo o mais recente registro de publicação datado em 16 de outubro de 2023. Ressalta-se a importância de seguir as orientações relativas à compatibilidade com o protocolo FTP. Para análises específicas dos dados da RAIS, a utilização do software R é aconselhada.

### **6.4. Receita Federal Dados Abertos**

A fonte de dados da Receita Federal disponibiliza informações sob diversas categorias, incluindo Arrecadação, Benefícios e Renúncias Fiscais, Cadastros, Carga Tributária, Comércio Exterior, Contencioso Administrativo, Convênios, Créditos Ativos, Distribuição de Renda, Fiscalização, Grandes Números do IRPF, Mercadorias Apreendidas, Órgãos e Municípios, Parcelamentos, ReceitaData, Restituição e Ressarcimento. Esses dados são representados em formato texto e podem ser processados por máquina, sendo disponibilizados sob licença aberta que permite sua livre utilização, consumo ou cruzamento. O Plano de Dados Abertos (PDA) do Ministério da Fazenda, agora Ministério da Economia, formaliza a estratégia de divulgação, sendo a última edição aprovada para o biênio 2023-2025, apresentando 10 novos conjuntos de dados a serem abertos. O relatório final do PDA/ME para o biênio 2021-2022 destaca a abertura de 42 bases de dados.

### **6.5. Dados Abertos MEC**

A fonte de dados proveniente do Ministério da Educação disponibiliza informações sobre diversos programas educacionais no Brasil. Os conjuntos de dados estão disponíveis nos formatos CSV e XML, com conjuntos específicos abrangendo anos diferentes. Por exemplo, o conjunto Escolas com plano de atendimento aprovado no Programa Mais Educação abrange os anos de 2014 a 2019, fornecendo detalhes sobre o

número de escolas, alunos, e valores recebidos. A frequência de atualização é anual. Outras fontes, como ProUni, Pronatec, e FIES, apresentam informações semelhantes, com detalhes sobre bolsas de estudo, instituições de ensino superior e financiamento estudantil. O Pronatec, em particular, fornece atualizações mensais, bimestrais, semestrais e anuais, dependendo do conjunto de dados específico. A cobertura geográfica é nacional, e as fontes abrangem diferentes períodos, com frequência de atualização anual, semestral ou específica para determinados eventos ou processos seletivos.

O conjunto de dados Escolas com plano de atendimento aprovado no Programa Mais Educação apresenta informações sobre o número de escolas municipais e estaduais que tiveram seus planos de atendimento do Programa Mais Educação aprovados. Os dados estão disponíveis nos formatos CSV e XML, abrangendo os anos de 2014 a 2019. O conjunto fornece detalhes, como a quantidade de alunos por escola, valores totais recebidos por adesão e a cobertura geográfica inclui todos os municípios brasileiros com beneficiários do Programa Mais Educação. A frequência de atualização é anual.

A fonte de dados do Programa Universidade para Todos (ProUni) oferece informações detalhadas sobre as bolsas de estudo integrais e parciais concedidas em instituições privadas de ensino superior. Os dados, disponíveis nos formatos CSV, abrangem o período de 2010 a 2016. Eles são segmentados por região, unidade federativa, município, instituição de ensino superior, curso, modalidade de ensino, turno e tipo de bolsa. A frequência de atualização é anual.

A fonte de dados do Programa Nacional de Acesso ao Ensino Técnico e Emprego (Pronatec) fornece uma lista abrangente de instituições da Rede Federal de Educação Profissional, Científica e Tecnológica. Os dados incluem detalhes como nome, município, data de autorização de funcionamento, quantidade de matrículas, novas matrículas e concluintes por iniciativa do Pronatec. A fonte abrange os anos de 2011 a 2016, com atualizações mensais, bimestrais, semestrais e anuais, dependendo do conjunto de dados específico.

A Plataforma Nilo Peçanha (PNP) é um ambiente virtual que coleta, valida e dissemina estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica. O conjunto de dados inclui informações como microdados de matrículas, servidores e eficiência acadêmica. A cobertura temporal é até o ano de 2021, com atualizações em outubro de 2022. Os dados estão disponíveis no formato CSV.

A fonte de dados do Fundo de Financiamento Estudantil (FIES) disponibiliza informações sobre os valores das semestralidades dos cursos ofertados por mantenedoras de instituições de ensino superior. Os dados abrangem o segundo semestre de 2019, com relatórios de resultados e inscrições até 2021. A cobertura geográfica inclui todos os municípios brasileiros com instituições participantes e candidatos inscritos no FIES. A frequência de atualização é anual.

A fonte de dados do Sistema de Seleção Unificada (SISU) disponibiliza dados relacionados às inscrições realizadas nos processos seletivos do SISU no formato CSV.

## **6.6. Dados Abertos INEP**

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é uma autarquia federal brasileira que desempenha um papel central na coleta e análise de dados educacionais. Os microdados fornecidos pelo INEP são fundamentais para compreender a complexidade do sistema educacional do Brasil em seus vários níveis. Incluindo informações detalhadas sobre escolas de Educação Básica, alfabetização nas séries iniciais e aspectos da educação superior, estes dados são cruciais para análises e avaliações profundas. Eles permitem que pesquisadores e formuladores de políticas avaliem a distribuição de recursos, a qualidade do ensino e identifiquem áreas que necessitam de melhorias. Esses dados também são essenciais no desenvolvimento de políticas educacionais eficazes, orientando decisões que impactam o progresso educacional, social e econômico do país.

## **6.7. Open Data SUS**

O Ministério da Saúde disponibiliza fontes de dados por meio do OPENDATASUS Estatísticas, composta por 31 conjuntos de dados relacionados à saúde no Brasil. Dentre esses conjuntos, destacam-se o Registro de Ocupação Hospitalar COVID-19, gerenciado pela Secretaria de Atenção Especializada em Saúde (SAES), que acompanha as internações durante a pandemia. Além disso, a vigilância da Síndrome Gripal é efetuada pela Secretaria de Vigilância em Saúde e Ambiente (SVSA), com notificações disponíveis para 2020, 2021 e 2022. Outros conjuntos incluem dados sobre a Campanha Nacional de Vacinação contra Covid-19, Saúde Indígena, Hospitais e Leitos, SRAG (Síndrome Respiratória Aguda Grave) em diferentes anos, SISAGUA relacionado à água, e informações sobre Unidades Básicas de Saúde (UBS) e Cadastro Nacional de

Estabelecimentos de Saúde (CNES). Os formatos dos conjuntos variam entre PDF, CSV, API, ODT e ZIP.

## **6.8. Conjunto de dados de Códigos Postais Mundiais**

A fonte de dados consiste em um banco de dados global contendo Códigos Postais de países de todo o mundo. Os conjuntos de dados representam cada país e incluem informações como Número do Código Postal, Nome do Local, Geolocalização, Precisão da Geolocalização e Código/Nome do Administrador. Os dados são disponibilizados em formato de string e podem ser acessados por meio de APIs REST ou GraphQL, permitindo fácil integração em várias aplicações. O banco de dados é licenciado sob a Creative Commons Attribution 4.0 License. Os conjuntos de dados individuais variam em tamanho, refletindo a diversidade de países incluídos. A última atualização foi registrada em 26 de fevereiro de 2020, e o serviço oferece suporte a consultas específicas, como listar Códigos Postais na Dinamarca ou contar todos os Códigos Postais na Suíça. Os desenvolvedores podem utilizar o serviço gratuitamente até 10 mil solicitações por mês e têm a flexibilidade de clonar e modificar o banco de dados conforme necessário.

## 7. Arquitetura da Solução

A arquitetura do sistema se refere às decisões que definem a estrutura e organização dos componentes que constituem a aplicação. Responsável por garantir que a aplicação seja escalável e segura. Abaixo é possível visualizar a primeira versão proposta da arquitetura da aplicação com suas camadas e um explicativo sobre cada módulo, posteriormente será apresentada a arquitetura definitiva, de forma mais detalhada sobre a parte técnica utilizada.

### 7.1. Arquitetura Versão Preliminar

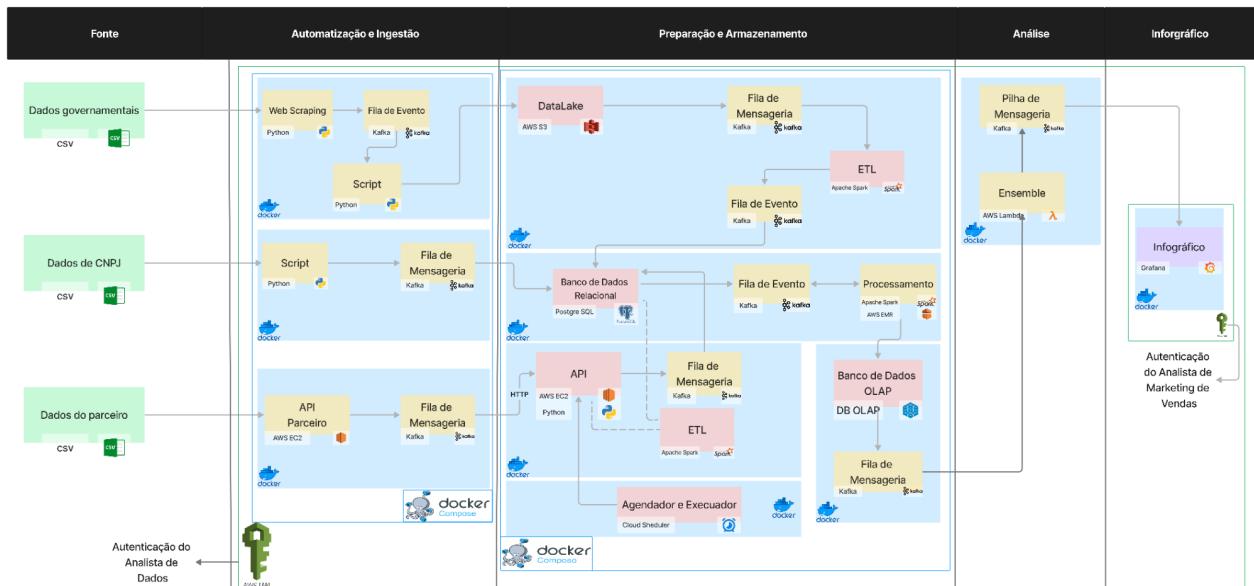


Figura 11: Arquitetura preliminar desenvolvida na Sprint 2.

Fonte: Elaboração própria

A arquitetura da solução foi desenvolvida para garantir que os dados sejam armazenados em um formato agnóstico à nuvem, permitindo que a solução seja migrada para qualquer provedor de serviços em nuvem, sem ficar restrita a uma única plataforma.

A segurança da aplicação desempenha um papel fundamental, e a arquitetura utiliza o AWS Identity and Access Management (IAM) para gerenciar o controle de acesso com base em funções de usuário, garantindo que apenas os usuários autorizados tenham acesso aos dados e recursos da aplicação.

## **7.1.1 Identificação dos dados de entrada e saída**

Nesse momento é necessário visualizar como serão as entradas e saídas de cada módulo da arquitetura, para melhor identificar as necessidades de cada etapa. Isso pode incluir dados de diferentes fontes, como bancos de dados, arquivos CSV, APIs, entre outros.

### **7.1.1.1. Dados de Entrada**

No geral, o pipeline recebe dados brutos de várias fontes, incluindo arquivos CSV de fontes governamentais, dados de CNPJ e informações de parceiros externos via API de consulta.

### **7.1.1.2. Dados de Saída**

Após o processamento, os dados são direcionados para várias saídas. Isso inclui o armazenamento em bancos de dados relacionais (PostgreSQL), bancos de dados OLAP e a plataforma Grafana para visualização de dados. A implementação de filas de mensagens e eventos em várias etapas do processo permite a transferência eficiente de dados entre os módulos.

Uma visão completa de como os dados fluem por todo o pipeline, desde a coleta até a análise e a visualização final. Cada módulo desempenha um papel importante na preparação e transformação dos dados para alcançar os objetivos de negócios.

## **7.1.2. Especificação de Entrada e Saída de cada módulo**

### **Módulo Fonte**

**Entrada:** Dados brutos de fontes governamentais (disponíveis via site e arquivos), dados de CNPJ (arquivos CSV) e informações dos parceiros externos via API de consulta.

**Saída:** Os dados são preparados e estruturados para serem transmitidos aos módulos subsequentes.

### **Módulo Automação e Ingestão:**

**Entrada:** Dados estruturados da etapa anterior, incluindo informações governamentais, dados de CNPJ e dados dos parceiros. Além disso, recebe constantemente novos dados conforme eles se tornam disponíveis nas fontes.<br>

**Saída:** Os dados processados são enviados para filas de eventos e filas de mensagens (Kafka) para transferência eficiente entre módulos, garantindo que as informações estejam prontas para serem consumidas e processadas na próxima etapa.

## **Módulo Preparação e Armazenamento:**

Entrada: Recebe os dados das filas de eventos e mensagens, que incluem informações governamentais, dados de CNPJ e dados dos parceiros. Integra informações recém-coletadas com os dados já existentes no DataLake AWS S3.<br>

Saída: Os dados são processados, limpos e transformados por meio de operações ETL (Extração, Transformação, Carregamento) usando o Apache Spark. Em seguida, os dados são armazenados em um DataLake no AWS S3, garantindo escalabilidade e armazenamento confiável. Os dados processados são inseridos em um Banco de Dados Relacional (PostgreSQL) e em um Banco de Dados OLAP otimizado para consultas analíticas.

## **Módulo Análise:**

Entrada: Recebe os dados processados armazenados no Banco de Dados Relacional (PostgreSQL) e no Banco de Dados OLAP. Além disso, recebe informações de eventos da fila.<br>

Saída: Executa um processo chamado Ensemble no AWS Lambda, que combina os dados e gera análises estatísticas e tendências. Os resultados são enviados para a plataforma de Infográficos (Grafana) para a criação de dashboards interativos e relatórios.

## **Módulo Infográfico:**

Entrada: Recebe os resultados do processo Ensemble e os dados armazenados no Banco de Dados OLAP.<br>

Saída: Cria e exibe infográficos interativos usando a plataforma de Infográficos (Grafana), permitindo aos analistas de marketing e vendas visualizar os insights e tendências de negócios de forma clara e concisa.

## 7.2. Arquitetura Final

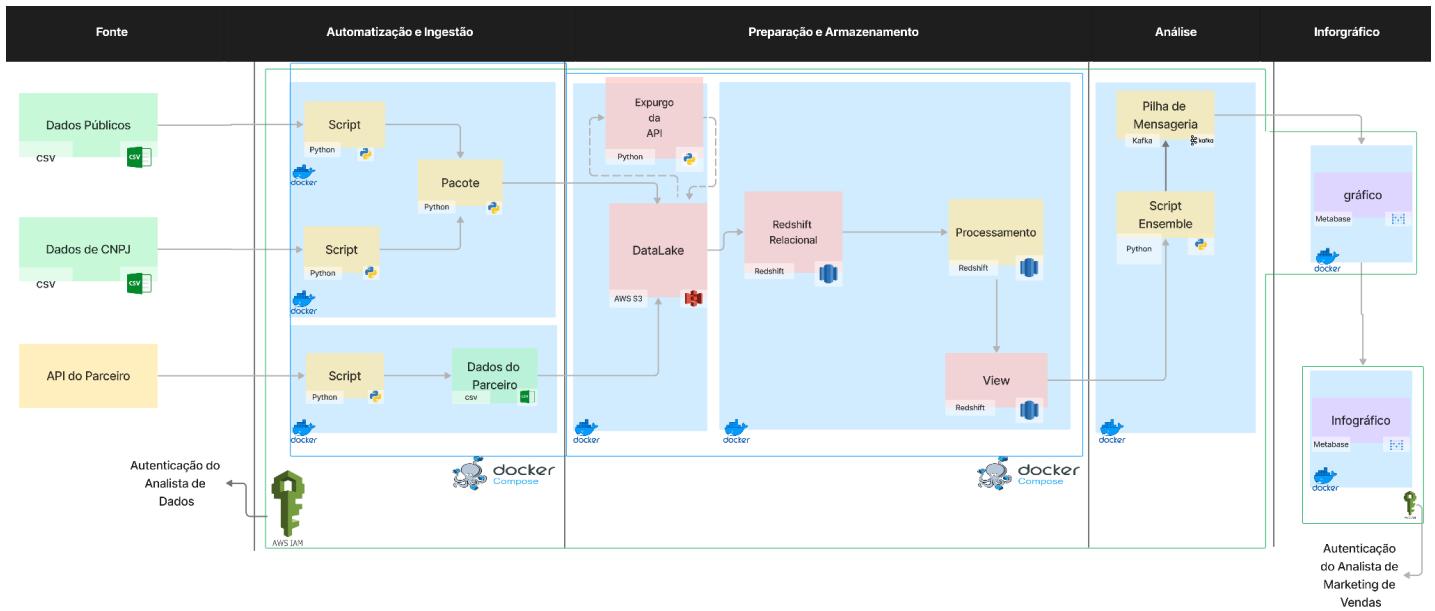


Figura 12: Arquitetura final do projeto

Fonte: Autores

A arquitetura final para a solução Big Data possui cinco etapas, iniciando com a captação de dados de diferentes fontes. Em seguida, automatiza e direciona os dados para processamento e armazenamento em AWS S3, com destaque para a remoção de dados sensíveis. Posteriormente, a análise dos dados por meio de scripts Python e o encaminhamento para criação de gráficos usando Metabase, permitindo a geração de infográficos. Cada etapa é descrita a seguir:

### 7.2.1. Fonte dos Dados

Foram mapeadas três fontes principais de dados, dando início ao processo:

- **Dados Públicos (CSV)**
- **Dados de CNPJ (CSV)**
- **Dados provenientes da API do parceiro de projeto**

Esta é a etapa para a identificação das origens dos dados a serem processados e analisados.

### 7.2.2. Automatização e Ingestão dos Dados

Ferramentas Utilizadas:

- **AWS IAM:** Fornece autenticação e controle de acesso.
- **Python (Script com Docker):** Automatiza a ingestão dos dados.

Nesta etapa, a automação e a correta ingestão dos dados são asseguradas por meio da autenticação com AWS IAM, restringindo o acesso aos analistas de dados autorizados, e as abordagens variam de acordo com a fonte dos dados:

- **Dados Públicos e de CNPJ:** São direcionados para um script Python executado em um ambiente Docker, encaminhando-os para um pacote específico.
- **Dados da API do parceiro de projeto:** São processados por um script Python separado, também utilizando um ambiente Docker específico para a construção de um arquivo (CSV).

### 7.2.3. Preparação e Armazenamento dos Dados

Ferramentas Utilizadas:

**AWS S3 (Data Lake):** Armazenamento escalável e durável.

**AWS Redshift:** Armazenamento relacional para processamento de dados.

**Docker:** Ambientes de execução isolados.

Os dados são recebidos e armazenados no Data Lake (AWS S3). Destaca-se que os dados da API do parceiro de projeto passam por um processo de expurgo para remover informações sensíveis. Em seguida, os dados são direcionados ao Redshift Relacional da AWS, onde são processados e permitem a criação de Views. Todo o processo ocorre em um ambiente Docker, garantindo segurança e isolamento.

### 7.2.4. Análise dos Dados

Ferramentas Utilizadas:

- **Python (Script Ensemble):** Coleta e processamento de informações.
- **Kafka:** Pilha de mensageria para transferência de dados em tempo real.

Nesta fase, um script em Python de método ensemble coleta e processa as informações provenientes das etapas anteriores. Os dados são então encaminhados para uma pilha de mensageria utilizando o serviço Kafka, garantindo a transferência em tempo real para análise.

### 7.2.5. Criação de Infográficos

Ferramentas Utilizadas:

- **Metabase**: Ferramenta de criação de gráficos e visualizações.
- **IAM (AWS)**: Controle de acesso gerenciado da AWS.

Nesta última etapa, a ferramenta Metabase é utilizada para criar gráficos e visualizações baseados nas informações processadas anteriormente. Adicionalmente, um componente dedicado à criação de infográficos, associado à autenticação IAM específica para analistas de marketing e vendas, utiliza os gráficos gerados para a composição de infográficos informativos.

### 7.2.6. Principais Diferenças

Enquanto a arquitetura final faz uso do Redshift Relacional da AWS e Metabase, a primeira versão utiliza Apache Spark, PostgreSQL, Grafana e AWS Lambda para processamento, armazenamento e análise.

Na arquitetura final, a análise é realizada via script Python e o serviço Kafka é utilizado para mensageria, enquanto na primeira versão, análises estatísticas e tendências são feitas via AWS Lambda e os infográficos são criados diretamente no Grafana.

## 7.3. Detalhes Técnicos do Pipeline de dados na AWS

Seção para documentar as estratégias adotadas, explicando o uso de serviços da Amazon, como o S3, Redshift, de forma a gerenciar o processamento e visualização de dados.

### **7.3.1. Pacote de Tratamento e Carga de dados Públicos na AWS S3**

Este pacote oferece funcionalidades para realizar tratamento superficial de dados e carregar os resultados para um bucket na AWS S3. O módulo incluído possibilita a manipulação de dados nulos, a conversão dos dados para um formato tabular específico e a renomeação de colunas para melhor entendimento.

#### **7.3.1.1. Funcionalidades**

- **Tratamento de Nulos:** Limpeza e imputação de valores em campos nulos para manter a integridade dos dados.
- **Conversão de Estrutura:** Transformação dos dados para um formato de tabela padronizado, facilitando operações subsequentes de análise e processamento.
- **Renomeação de Colunas:** Atualização dos nomes das colunas para termos mais descriptivos, visando uma visualização e compreensão mais clara dos dados.

#### **7.3.1.2. Uso**

Após aplicar o tratamento de dados e enviar para a AWS S3, utilize a função `load_s3` da classe, preenchendo os parâmetros necessários. A função realizará o upload dos dados tratados para o bucket especificado na AWS S3.

##### **7.3.1.2.1. Pré-Requisitos**

Dados de entrada devem estar em um formato CSV para que a classe possa processar.

Credenciais de acesso à AWS devem estar configuradas corretamente.

##### **7.3.1.2.2. Funções**

`tratar_nulos(self)`

Realiza a limpeza de campos nulos nos dados, utilizando a estratégia de imputação definida.

`converter_arquivo(self, type)`

Converte o arquivo de dados para o formato de tabela específica. O parâmetro type, está sendo utilizado para criar uma coluna que descreva a origem dos dados, facilitando a futura união de tabelas e permitindo diferenciá-las, se necessário.

```
substituicao_nomenclaturas(self, dicionario_de_substituicao)
```

Para renomear as colunas dos dados, aplique um dicionário de substituição. Para usar essa função, simplesmente insira como argumento o dicionário a ser utilizado para a renomeação, que geralmente tem o mesmo nome da tabela. Este pode ser facilmente encontrado no arquivo localizado no mesmo diretório, chamado 'dic\_pof'.

```
load_s3(self, bucket_name, object_name, aws_access_key_id,  
aws_secret_access_key, token)
```

Carrega os dados tratados para um bucket na AWS S3. Requer todos os parâmetros necessários para autenticação e especificação do destino dos dados.

### 7.4.1. Pacote de Tratamento e Carga da API do Parceiro na AWS S3

Este pacote foi desenvolvido para integrar dados provenientes da API do Parceiro, tratá-los e carregá-los em um bucket da AWS S3. A funcionalidade principal é facilitar a automação do processo de extração de dados, seu tratamento preliminar e o armazenamento na nuvem, permitindo que usuários e sistemas downstream acessem os dados prontos para uso.

#### 7.4.1.1. Funcionalidades

**Consulta à API:** Permite a realização de consultas programáticas à API do Parceiro, obtendo dados recentes e relevantes.

**Tratamento de Dados:** Aplica um conjunto de procedimentos para garantir que os dados estejam limpos e estruturados corretamente antes do carregamento.

**Carga no S3:** Facilita o envio dos dados tratados para um bucket específico na AWS S3, tornando-os acessíveis para análise e aplicativos.

#### 7.4.1.2. Uso

O fluxo típico de uso deste pacote envolve a chamada da função de consulta à API do Parceiro, o tratamento dos dados recebidos e o subsequente carregamento no S3.

#### **7.4.1.2.1. Pré-Requisitos**

É necessário possuir as credenciais de acesso API do Parceiro.

As credenciais da AWS também devem estar configuradas para permitir o acesso ao serviço S3.

#### **7.4.1.2.2. Funções**

```
consulta_api(self, endpoint, parametros)
```

Realiza uma requisição GET ao endpoint especificado da API do Parceiro, utilizando os parâmetros fornecidos. Retorna os dados em formato JSON.

```
tratar_dados(self, dados_json)
```

Recebe os dados em formato JSON, aplica o tratamento necessário e os converte para um DataFrame do pandas, preparando-os para o carregamento no S3.

```
carregar_s3(self, bucket_name, object_name, dados_dataframe)
```

Carrega o DataFrame tratado no bucket da AWS S3 especificado, utilizando o nome do objeto fornecido para o arquivo resultante.

```
executar_carga(self)
```

Orquestra o processo de consulta à API, tratamento dos dados e carga no S3. Esta função chama internamente as funções consulta\_api, tratar\_dados e carregar\_s3.

#### **7.4.1.3. Exemplo de Uso**

```
# Instanciando a classe
api_parceiro = ConexaoAPIParceiro(api_key, secret_key, token)
# Executando a carga completa
api_parceiro.executar_carga()
```

## **7.5. Modelo de Regressão para prever o coeficiente do nível de "Pobreza Multidimensional Não Monetária" (IPM-NM)**

Este modelo de regressão desempenha um papel essencial na compreensão e análise da pobreza multidimensional, ultrapassando as considerações ligadas à renda financeira.

Além disso, a importância deste modelo transcende a mera previsão da pobreza monetária, estendendo-se à avaliação do consumo de produtos alimentícios. Ao incorporar variáveis multidimensionais, o modelo proporciona insights sobre as condições de vida e os fatores que contribuem para a pobreza em diversas dimensões, incluindo a alimentação.

### **7.5.1. Base de dados utilizadas**

Serão utilizadas duas bases consolidadas no ano de 2008 e 2009, e outra do ano de 2017 e 2018. A fim de unir as duas tabelas, serão concatenadas para buscar a diferença entre os dois períodos.

### **7.5.2. Dados para treinamento**

Estamos utilizando a função `train_test_split` da biblioteca scikit-learn para dividir seus dados em conjuntos de treinamento e teste. X são suas características e y é a variável de resposta.

- `test_size=0.2`

Significa que 20% dos dados serão usados como conjunto de teste, e o restante (80%) será usado como conjunto de treinamento.

- `random_state=42`

É uma semente para garantir reproduzibilidade. Isso significa que, se você executar o código novamente com a mesma semente, obterá a mesma divisão de treinamento/teste.

## **7.6. Infográfico - Relatório de Análise de Eficácia e Sugestões de Melhoria no Projeto**

Este relatório fornece uma análise detalhada e crítica dos infográficos desenvolvidos para o projeto da Integration Consulting, de Gerenciamento e Análise de Big Data. O objetivo central é avaliar a capacidade dos infográficos de representar o consumo por canal, região e categoria, buscando gerar insights valiosos para o setor de consultoria em um ambiente de Big Data.

### **7.6.1. Metodologia**

Adotamos uma metodologia para a síntese de dados complexos de Big Data em infográficos visuais, utilizando os recursos avançados da AWS. Esta abordagem envolveu a análise e transformação de dados em representações gráficas, incluindo ícones e ilustrações complementadas por textos, visando a uma compreensão e assimilação mais intuitiva das informações.

### **7.6.2. Análise dos Infográficos**

#### **Pontos Fortes**

Os infográficos, desenvolvidos com um enfoque específico na consumação por canal, região e categoria, destacam-se por sua clareza e facilidade de acesso. Eles proporcionam insights profundos sobre o consumo, as características dos clientes e padrões de vendas. A diversidade em tipos de gráficos, uma paleta de cores variada e a implementação de filtros são recursos estratégicos que enriquecem a narrativa visual, tornando os dados mais atraentes e de fácil interpretação.

#### **Pontos Fracos**

Apesar dos avanços, enfrentamos desafios significativos devido ao volume de dados e arquivos, levando a dificuldades na interpretação para alguns usuários, particularmente em relação à complexidade das legendas e distinção de cores. Reconhecemos também a importância de um pré-processamento de dados meticuloso

para a construção eficaz de infográficos, uma tarefa que pode ser intensiva em termos de tempo e esforço.

### **7.6.3. Descrição dos Gráficos**

Os gráficos são detalhadamente descritos, destacando sua estrutura, tipologia e representações visuais. Eles foram criados considerando as etapas de ingestão, preparação e análise estatística de dados na infraestrutura da AWS, visando maximizar a clareza e o impacto das informações apresentadas.

### **7.6.4. Pontos de Melhoria**

Propomos melhorias como a inclusão de diferenciações visuais mais acentuadas, destacando informações específicas como alimentos por cidade e cor. A análise detalhada de CNPJs e o uso de dados reais são sugeridos para aumentar a precisão e aplicabilidade dos infográficos. Estas melhorias visam aprimorar a experiência do usuário, tornando os infográficos mais interativos, acessíveis e informativos.

## **8. Análise de Custos e Impacto da Solução**

### **8.1. Impactos Esperado da Solução**

Uma solução bem-sucedida para este problema terá vários impactos positivos:

*Tomada de Decisão Informada:* A solução fornecerá informações valiosas que ajudarão a Integration e seus clientes a tomar decisões estratégicas bem fundamentadas em relação à distribuição, identificando oportunidades de crescimento e otimização.

*Eficiência Operacional:* A integração automatizada de dados e a análise estatística reduzirão a dependência de processos manuais demorados, economizando tempo e recursos.

*Competitividade Aprimorada:* A capacidade de analisar dados de forma eficaz permitirá à Integration e a seus clientes se manterem competitivos no mercado de distribuição, adaptando-se às mudanças nas preferências do consumidor e nas condições de mercado.

*Comunicação Visual Clara:* O infográfico facilitará a comunicação de informações complexas de maneira acessível, tornando mais fácil para todas as partes interessadas entender e usar os resultados da análise.

### **8.2. Análise de Custos**

O documento de análise financeira apresenta os aspectos monetários do projeto, com a projeção dos custos, investimentos esperados ao longo de um período específico, permitindo que stakeholders e investidores compreendam a viabilidade financeira e os possíveis retornos sobre o investimento.

#### **8.2.1. Serviços da Amazon Services Utilizados:**

##### **8.2.1.1. Amazon Redshift:**

Serviço de armazenamento e análise de dados de alta performance, desenvolvido para processar grandes volumes de informações. Baseado na tecnologia de data warehousing, o Redshift permite a criação e gerenciamento de data warehouses em escala, proporcionando aos usuários a capacidade de executar consultas e análises nos conjuntos de dados.

#### **8.2.1.2. S3 (S3 Standard e Data Transfer):**

Serviço de armazenamento em nuvem altamente durável, projetado para recuperar dados em casos de falhas. Por ser uma opção padrão para armazenamento, é ideal para uma variedade de casos de uso, favorecendo a disponibilidade e desempenho, gerenciando a transferência de dados para dentro e para fora do serviço, permitindo o tráfego de informações entre diferentes serviços AWS.

#### **8.2.1.3. AWS Lambda:**

Serviço de computação serverless que permite a execução de código de forma automática em resposta a eventos, e triggers, possibilitando o processamento em tempo real para automação de tarefas e construção de aplicativos sem servidor.

#### **8.2.1.4. Amazon EC2 (Elastic Compute Cloud):**

Oferece a capacidade de computação escalável na nuvem, para o provisionamento de servidores virtuais configuráveis para executar aplicativos possibilitando adaptar a infraestrutura de computação de acordo com as necessidades de processamento, memória, armazenamento e outros recursos.

### **8.2.2. Custos Projetados para o Primeiro Ano:**

#### **8.2.2.1. Investimento Inicial:**

Custos iniciais do projeto: \$0 USD.

#### **8.2.2.2. Custo das ferramentas:**

- Amazon Redshift:

Custo mensal: \$527.04 USD;

Custo em 12 meses: \$6324.48 USD;

Serviço: RPU base (16), Tempo de execução diário esperado (3 horas).

- S3 (S3 Standard e Data Transfer):

Custo mensal: \$0.57 USD (S3 Standard), \$0.22 USD (Data Transfer);

Custo em 12 meses: \$9.48 USD;

Serviço: S3 Standard, Data Transfer.

- AWS Lambda:

Custo mensal: \$0 USD;

Custo em 12 meses: \$0 USD;

Serviço: Arquitetura (x86), Quantidade de armazenamento temporário alocada (10 GB), Modo de invocação (Em buffer), Número de solicitações (50 por dia).

- Amazon EC2:

Custo mensal: \$118.26 USD;

Custo em 12 meses: \$1419.12 USD;

Serviço: Locação (Instâncias compartilhadas), Sistema operacional (Linux), Carga de trabalho (Consistent, Número de instâncias: 10), Instância do EC2 avançada (t3a.medium), Pricing strategy (Amazon EC2 Instance Savings Plans 3yr No Upfront), Habilitar monitoramento (desabilitada)

#### **8.2.2.3. Custo Total:**

Por Mês: \$646.09 USD

Por 12 Meses: \$7,753.08 USD

Data da cotação dos valores: 7 de dezembro de 2023

Link para a cotação:

<https://calculator.aws/#/estimate?id=0fad95fbb78847899f1061c57e8a9b536733e140>

### **8.2.3. Observação**

Para manter a consistência na alimentação e armazenamento de dados no cubo OLAP, das quatro ferramentas da AWS delineadas na arquitetura do projeto: Amazon Redshift, Amazon S3, AWS Lambda e Amazon EC2, a projeção dos custos revela que o principal investimento será direcionado ao armazenamento no serviço Redshift da AWS, com um custo mensal estimado de \$527.04 USD. Por outro lado, os serviços Lambda e S3 da Amazon totalizam um valor mensal de \$0.79 USD combinados. Por fim, as instâncias da Amazon EC2 têm um custo mensal estimado de \$118.26 USD, e dessa forma, a projeção total dos custos do projeto, considerando esses serviços da AWS, \$646.09 USD ao longo de um ano.

## **9. Prototipação em Baixa Fidelidade com Wireframes**

Antes da construção de um sistema de dashboard para a visualização de dados e infográficos, por exemplo, é importante a utilização de uma ferramenta de prototipação que possibilite ajustes ágeis que não representam grande impacto no tempo e retrabalho de mudanças, que afetariam a qualidade da entrega do projeto. Com isso, protótipos feitos com wireframes são úteis para usuários do sistema apontarem possíveis problemas e apresentarem feedbacks que podem incorporar na construção posterior de design do sistema. Nesse contexto, os protótipos criados atuam como um guia esquemático que delinea a distribuição de elementos-chave como caixas de texto, botões, imagens, mas sem aprofundamentos gráficos.

### **9.1. Wireframes Desenvolvidos**

Duas páginas foram criadas via wireframe, projetadas de acordo com o impacto mapeado na experiência e necessidade do usuário. Na primeira página, destinada ao acesso inicial de um usuário autenticado, foram priorizados elementos que fornecem uma visão imediata de informações críticas e detalhadas. O design desta página é orientado pela estratégia de destacar visualmente o geomapa de dados de vendas, gráficos de vendas por canais e consumo por categoria, além de rankings para dados de vendas e canais. Esses elementos prioritários foram escolhidos para orientar o usuário rapidamente às métricas-chave, facilitando a tomada de decisões estratégicas.

Por outro lado, a segunda página, desenvolvida para complementar a primeira, atuando como um suporte, oferece uma visualização mais abrangente. Com três filtros distintos, como região, canal e categoria no elemento principal, um gráfico de percentual dos dados usados na ingestão e uma tabela dos dados, essa página foi concebida para aprofundar a análise e proporcionar uma perspectiva que assegure a integridade dos dados.

Juntas, essas páginas formam uma experiência que equilibra o acesso de informações prioritárias com profundidade analítica e também a validação necessária para a compreensão dos dados presentes nos infográficos.

Mais informações sobre a escolha da posição dos elementos e técnicas utilizadas e feedbacks utilizados para elaborar e ajustar os protótipos wireframe das páginas podem ser verificados no tópico 3. Técnicas Aplicadas. A seguir os protótipos desenvolvidos:

Visando melhor atender a necessidade do cliente, também foi feita a versão mobile do wireframe das páginas:

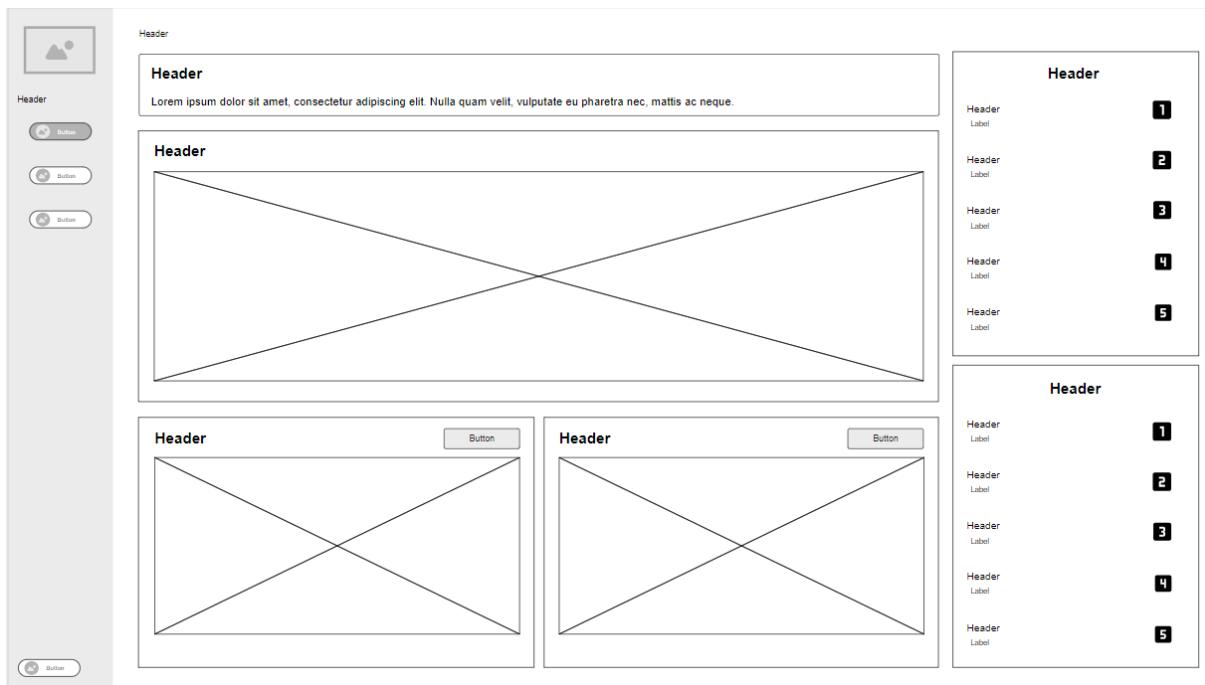


Figura 13: Wireframe Web 1. Fonte: Elaboração própria

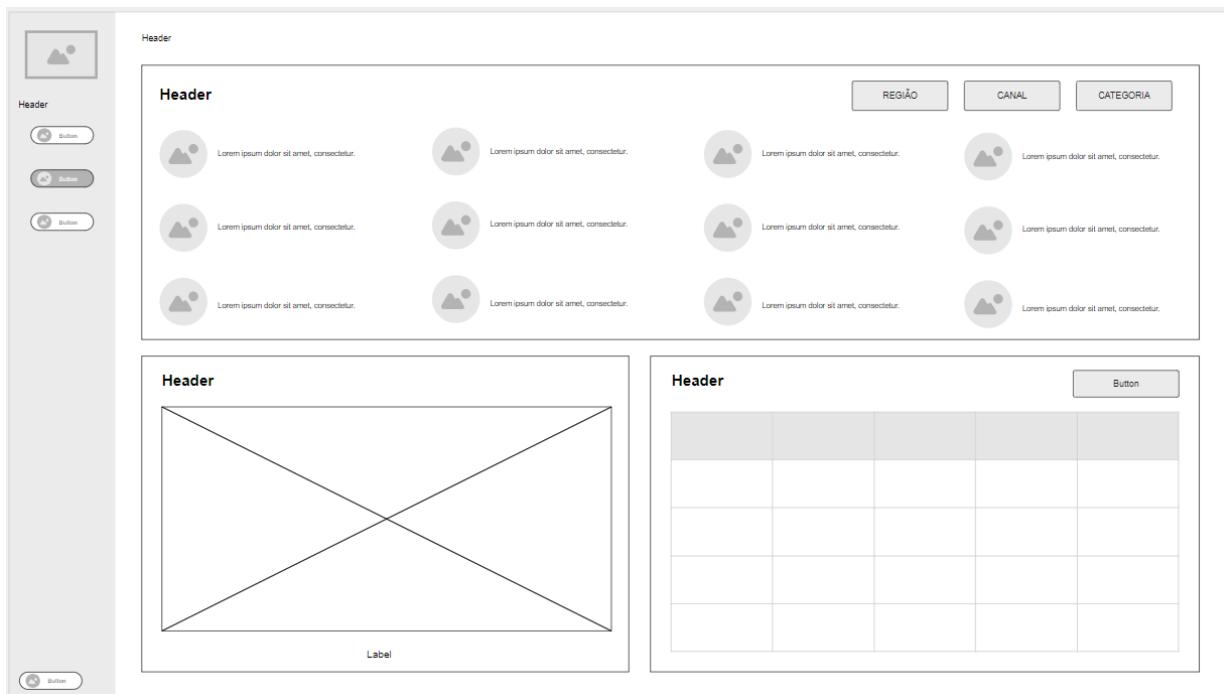


Figura 14: Wireframe Web 2. Fonte: Elaboração própria

Visando melhor atender a necessidade do cliente, também foi feita a versão mobile do wireframe das páginas:

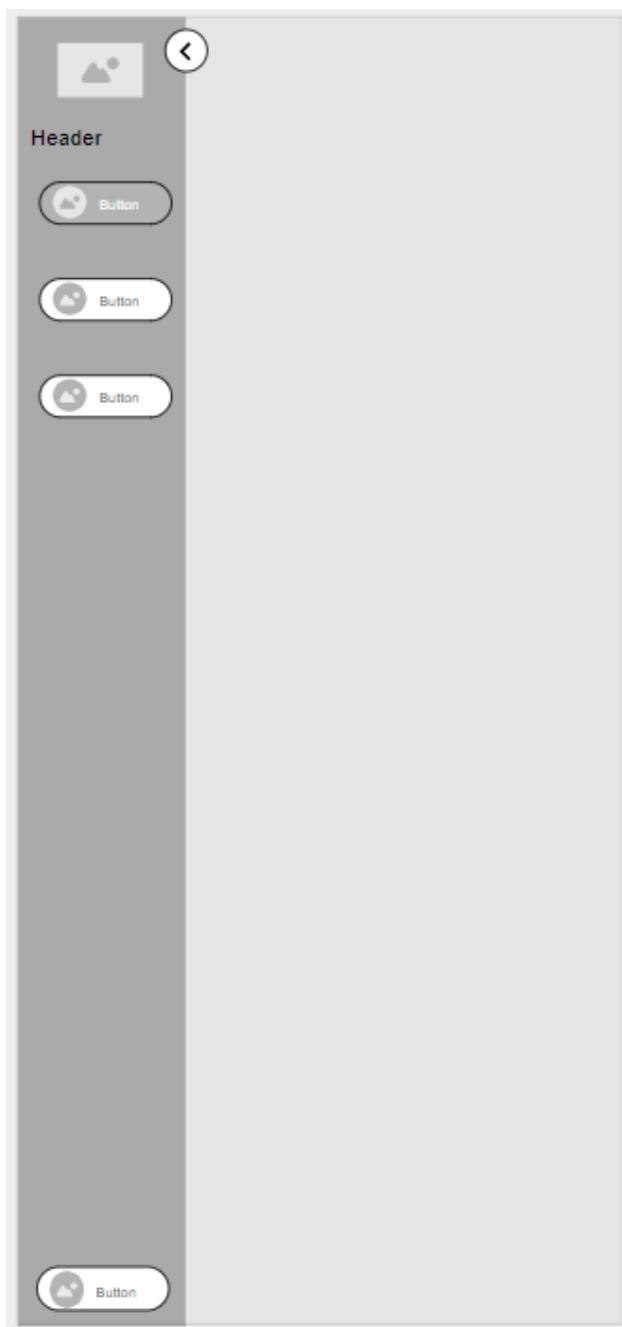


Figura 15: Wireframe Mobile 1. Fonte: Elaboração própria

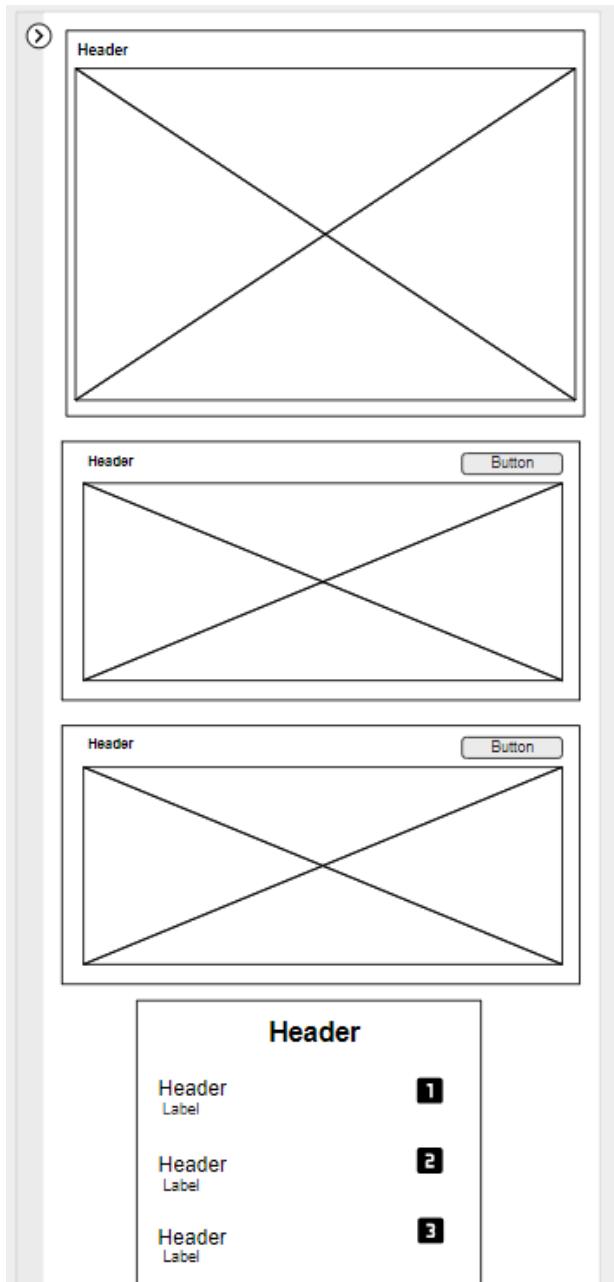


Figura 16: Wireframe Mobile 2. Fonte: Elaboração própria.

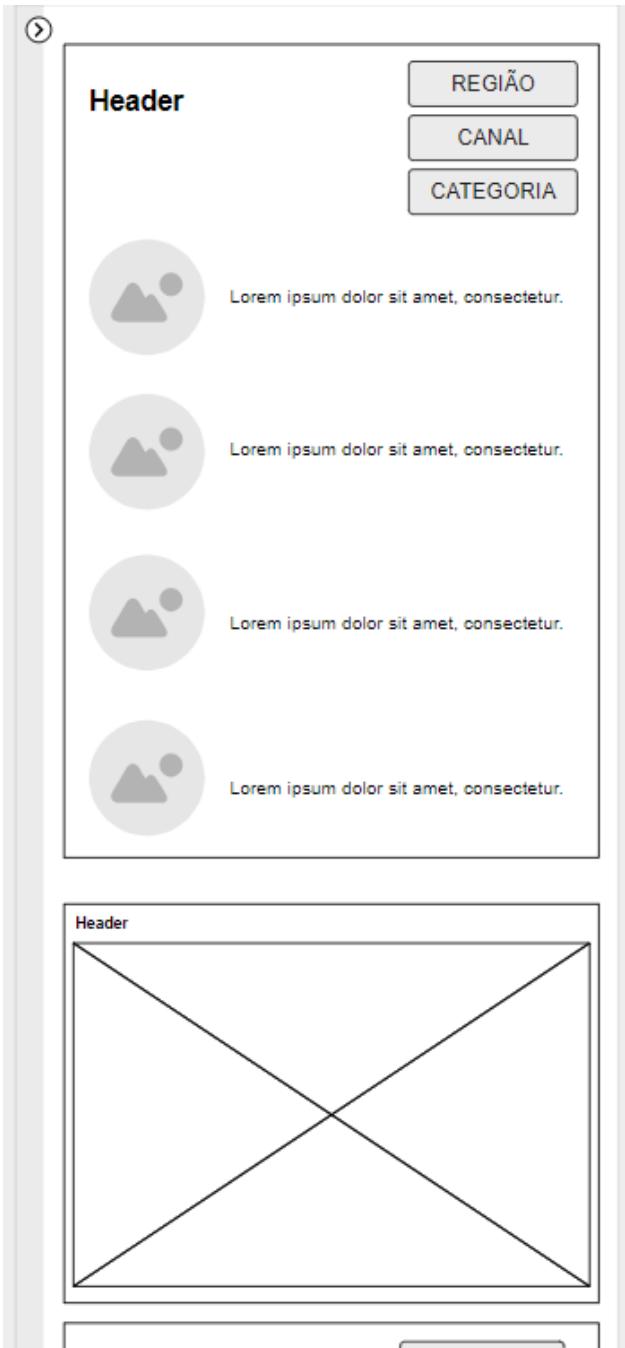


Figura 17: Wireframe Mobile 3. Fonte: Elaboração própria

## 9.2. Técnicas Aplicadas

Os wireframes desenvolvidos tiveram uma abordagem centrada na usabilidade e na interpretação dos dados, onde a escolha do design foi fundamentada em técnicas e feedbacks recebidos, aplicados de forma diferente a cada elemento, de acordo com a sua importância e dependência de outros dados, que serão descritos a seguir:

### **9.2.1. Ênfase na Prioridade dos Dados**

O espaço destinado a dados que devem estar presentes assim que o usuário acessar o sistema deve representar o primeiro elemento que usuário irá naturalmente olhar, seguindo o princípio da visão em "Z" das páginas web, justificando o geomapa com dados de vendas como elemento prioritário, refletindo a importância estratégica dessas informações para o usuário do sistema. A partir deste elemento, foi pontuada a necessidade da função de granularidade dos elementos mostrados, apresentando também uma flexibilidade por filtros que podem ser implementados no geomapa real implementado no sistema.

### **9.2.2. Organização de Informações para Análise Cruzada**

A análise de dados pode se beneficiar com a criação de um dashboard, dispondo elementos que podem se complementar, um ao lado do outro. Os espaços menores adjacentes destinados aos gráficos de vendas por canais e consumo por categoria e os rankings de vendas por categoria e canal estão dispostos em uma organização que facilita a comparação e análise cruzada dessas métricas. Um feedback recebido sobre ambos grupos de elementos foi aumentar a flexibilidade do que é mostrado, adicionando a possibilidade de trocar as variáveis que são mostradas, já adiantando próximos passos ao evoluir o protótipo feito em wireframe a um nível de maior fidelidade.

### **9.2.3. Dados comparativos por Rankeamento de Variáveis**

A inclusão de dois rankings proporcionam uma rápida visualização por desempenho em termos de vendas e canais, priorizando informações para a verificação dos elementos que apresentam maior performance em relação aos demais, trazendo possíveis pontos de atenção que devem ser priorizados a partir de uma grandeza de tamanho, volume ou quantidade. A adição de filtros para trocar a variável utilizada para a criação do ranking foi descrita anteriormente.

### **9.2.4. Visualização Total com Opção de Filtros**

Com o segundo wireframe representando a segunda página do sistema, a prioridade dos dados visualizados é reduzida, mas devem complementar os dados apresentados na primeira página. A necessidade de uma visão total de um tipo de elemento é representada como o primeiro elemento visível para o usuário na segunda página, com a visualização dos parceiros, que foi projetado considerando a necessidade

de explorar esses dados de maneira a verificar a situação desses elementos, que a partir dos principais filtros por região, canal e categoria, foram apontados como feedback para a personalização da visualização pelo usuário.

### **9.2.5. Status dos Dados Utilizados**

Ainda na segunda página, para assegurar o usuário dos dados utilizados para a criação dos elementos da primeira página é foi planejado um espaço que apresente dados sobre a integridade dos dados presentes nos infográficos gerados, e isso é representado com o percentual dos dados e um filtro dos dados que informam o usuário sobre os dados que foram usados na ingestão dos dados, promovendo maior confiabilidade e balizando decisões tomadas com a segurança dos dados que estão em uso.

### **8.2.6. Telas Menores e Gestos Típicos de Dispositivos Móveis**

No wireframe que representa a primeira página, por apresentar os elementos com maior prioridade de visualização, foi mantida uma configuração de manter os elementos mais relevantes de serem visualizados primeiro no topo da página, funcionando de maneira independente em relação aos demais e parear os elementos menores por dependência, em no máximo dois. Com isso, os gráficos que podem proporcionar uma análise cruzada permanecem um ao lado do outro em telas menores, bem como os dados de ranking.

Já na segunda página de wireframe, cada elemento deve estar posicionado na mesma coluna, um abaixo do outro, favorecendo a visualização de cada elemento de forma independente, pois eles não se complementam diretamente a ponto de proporcionar uma análise cruzada, como na primeira página.

### **9.2.7. Reduzir a Utilização de Texto:**

Foi utilizado um header de tamanho reduzido, considerando a importância de comunicar ao usuário, por exemplo, quando aqueles dados foram atualizados ou quando os infográficos presentes na página foram gerados, para o acompanhar as versões de gráficos gerados, trazendo maior confiabilidade nas análises e decisões tomadas. Pela relevância da informação, este header foi colocado acima de todos os gráficos na primeira página, podendo sofrer alterações de acordo com o avanço da prototipação em fidelidades mais altas.

É importante também destacar que todos os elementos do protótipo podem sofrer ajustes e alterações de acordo com a evolução do protótipo em níveis de maior fidelidade e com base em testes de usabilidade, bem como feedbacks do parceiro e usuário.

### **9.2.8. Incorporação de Feedbacks:**

Após a apresentação dos wireframes, foram recebidos feedbacks pelo parceiro e pelo professor, que estão sendo incorporados na evolução dos protótipos e irão ser demonstrados nas próximas entregas. Os feedbacks recebidos foram:

Sempre tentar cruzar dados de categoria e canal, para que o usuário possa ter uma visão mais completa dos dados;

Incluir um filtro de data para que o usuário possa selecionar o período de tempo que deseja visualizar;

Sobre a questão de região, foi dito que quanto mais granularidade, melhor, pois o usuário pode querer visualizar dados de uma região específica;

No painel, devemos incluir uma combinação de gráficos, tabelas e KPIs. Em vez do ranking, poderíamos usar alguns KPIs.

## **10. Análise de Impacto Ético**

Big Data tornou-se uma ferramenta fundamental em diversos setores, permitindo análises complexas e tomadas de decisão baseadas em dados. Seu uso abrange desde a personalização de serviços até avanços na medicina e na conservação ambiental. Contudo, junto com seus benefícios, surgem desafios éticos e sociais que precisam ser abordados.

Neste documento serão abordados os impactos deste projeto, destacando a sua influência em cinco dimensões, discutindo as formas escolhidas para mitigar os impactos negativos, e promover os aspectos positivos para a sociedade.

Dentro dos conjuntos de dados de Big Data podem estar inclusos informações sensíveis, que possibilitam a identificação, de forma direta ou indireta, do indivíduo a quem o dado pertence, expondo nome, residência e rotinas pessoais, e isso pode ser exposto de forma intencional, ou não (White et al., 2019). Isso envolve a coleta, armazenamento e o uso desses dados, provocando uma reação de implementar a regulamentação e governança desses dados, de forma a proteger os dados contra acesso

não autorizado e uso indevido, definindo também a quem o dado pertence e implicações legais sobre o uso, o armazenamento e a coleta dos dados.

Porém, segundo Tene e Polonetsky (2013), a proteção desses dados se torna cada vez mais difícil, a partir do momento em que estes dados são compartilhados entre diversas plataformas e empresas ao redor do mundo, carregando informações como saúde, finanças, consumo e o comportamento online dos indivíduos, provocando uma preocupação sobre a perda sobre o controle e o rastreamento do uso desses dados.

Por isso é importante destacar que, para o público em geral, além da segurança sobre os dados, é necessária a confiança sobre o armazenamento e uso desses dados, visto que diversas empresas e órgãos governamentais sofrem vazamentos dos dados, e o indivíduo pode pensar que o dado pertence a ele, quando na verdade o dado pode legalmente pertencer à empresa que o coletou, a depender do país e os termos de uso sobre a plataforma ou software em que o usuário inseriu os dados (White et al., 2019).

## 10.1. Privacidade e Proteção de Dados

Os desafios e implicações de privacidade e proteção de dados diferem em relação à fonte e o tipo de dado utilizado por um sistema, e devem considerar aspectos legais e práticas éticas para lidar com dados sensíveis e não sensíveis. Visando garantir a conformidade dos dados com regulamentações de privacidade e a categorização adequada, foi considerado o tratamento de acordo com a fonte dos dados utilizados:

Os Dados Públicos são predominantemente coletados e gerenciados pelo governo, não estando sujeitos a restrições de direitos autorais, patentes ou marcas registradas, onde restrições de privacidade, controle e segurança devem ser regulamentadas por estatutos, sendo que os dados públicos têm propensão para serem abertos, mas isso não inclui informações pessoais sob custódia estatal ou detalhes particulares de empresas ou organizações específicas (Possamai, 2016), podendo estes serem armazenados e utilizados de forma livre considerando a privacidade e a proteção da informação.

Da mesma forma, Dados de CNPJ, que envolverem informações associadas a registros de empresas, não são considerados dados pessoais para efeitos da Lei, segundo a LGPD, porém conferem um tipo de dado estratégico da empresa parceira,

sendo necessário assegurar a sua integridade, garantindo a segurança ao ser armazenado no banco de dados da solução.

Já os Dados provenientes da API do Parceiro configuram dados estritamente sensíveis, contendo dados confidenciais dos clientes da empresa de consultoria, requerendo uma abordagem específica em termos de segurança, privacidade e conformidade.

Uma regra de negócio envolve a remoção dessas informações após o consumo por ferramentas de análises de dados, reduzindo o risco de exposição ou acesso não autorizado, aumentando a confiança dos clientes sobre os serviços prestados pela empresa. Essa remoção é uma estratégia de tratamento realizada por scripts, definidos na arquitetura projetada, que garantem que as informações não permaneçam armazenadas após determinados ciclos do seu uso e consumo.

## **10.2. Equidade e Justiça**

Segundo Taylor (2017), a partir do avanço tecnológico, populações que antes eram digitalmente invisíveis passaram a ter dados relevantes para pesquisadores e formação de novas políticas públicas, porém esse avanço é limitado devido a dificuldade na equidade da utilização e tratamento dos dados, ou seja, mesmo com representação nas bases de dados, alguns grupos e comunidades seguem sendo prejudicados com dados desproporcionais e insuficientes.

Embora Big Data ofereça oportunidades, é crucial considerar e mitigar possíveis disparidades. Isso inclui garantir acesso igualitário à tecnologia e aos benefícios derivados dela, bem como identificar e corrigir vieses que poderiam resultar em tratamento injusto ou desigual entre grupos sociais.

As disparidades entre regiões e classes sociais frequentemente se refletem nos conjuntos de dados disponíveis. Essas diferenças podem surgir devido a variados níveis de acesso à tecnologia, infraestrutura digital e fatores socioeconômicos. Regiões ou comunidades com infraestrutura digital precária podem estar sub-representadas nos conjuntos de dados, levando a uma menor visibilidade nas análises de dados.

Uma medida aplicada para mitigar isso é implementar uma maior granularidade dos resultados gerados, que se refere à capacidade de analisar informações em níveis mais detalhados e específicos. Isso pode envolver a segmentação dos dados por características demográficas, geográficas ou outros fatores relevantes. Ao realizar análises mais granulares, é possível identificar padrões, necessidades e desafios específicos enfrentados por diferentes regiões, grupos étnicos, faixas etárias ou classes sociais, garantindo que as análises abordem as necessidades específicas de cada grupo ou região.

### **10.3. Transparência e Consentimento Informado**

Enfoque na importância da transparência na coleta, uso e compartilhamento de dados no contexto do Big Data. Explicação sobre o consentimento informado e seu papel na construção de relações de confiança com os usuários. Práticas recomendadas para garantir a transparência e obtenção adequada do consentimento.

Embora a transparência envolva a clareza e acessibilidade das informações fornecidas sobre como os dados são coletados, usados e compartilhados com a permissão concedida sobre o uso de seus dados, ferramentas baseadas em dados nem sempre podem ser completamente transparentes, dependendo da complexidade técnica da solução e os algoritmos utilizados (Varley-Winter e Shah, 2016).

Nesse sentido, a confiança passa a ser um pilar relevante para estabelecer um entendimento sobre as práticas de coleta, tratamento, armazenamento e uso dos dados, a partir de uma governança dos dados, de forma a assegurar que as partes envolvidas entendam os processos, dados necessários, e os dados utilizados no sistema e do serviço prestado, com políticas de privacidade detalhadas e cláusulas de consentimento explícitas.

### **10.4. Responsabilidade Social**

A responsabilidade social se refere ao compromisso das organizações em agir de maneira ética e sustentável, considerando não apenas o lucro, mas também os impactos

sociais, ambientais e comunitários de suas ações, minimizando impactos negativos e alinhando-se a metas de desenvolvimento sustentável.

É importante considerar os impactos mais amplos do uso de Big Data, englobando avaliar os efeitos do projeto sobre comunidades, práticas sustentáveis e buscar contribuir para metas de desenvolvimento social e ambiental, fazendo parte de uma responsabilidade social, podendo ser alinhada aos Objetivos de Desenvolvimento Sustentável (ODS) da ONU.

Mesmo com dados categorizados como não sensíveis, o uso de dados pode ter um efeito negativo sobre a sociedade, perpetuando a estratificação social, com a exclusão de indivíduos que, ao usarem softwares e participarem de atividades com registros online e cederem seus dados em troca do uso e participação, não recebem benefícios provenientes dos dados analisados. Empresas podem, por exemplo, precificar produtos e serviços no limite de pagamento de usuários, usando análise de dados para obter uma maior margem de lucro, porém essas mesmas empresas raramente estão dispostas a compartilhar a riqueza gerada com os dados dos indivíduos com os próprios indivíduos que geraram esses dados (Tene e Polonetsky, 2013).

Nesse sentido, considerando o setor da empresa e os dados coletados e ingeridos, o impacto na ODS de Fome Zero e Agricultura Sustentável pode ser considerável, promovendo ações que podem identificar e auxiliar comunidades carentes com vulnerabilidade alimentar. E ainda por promover uma nutrição mais saudável para essas comunidades, também promoveria positivamente o objetivo de Saúde de Qualidade, melhorando a segurança alimentar e reduzindo doenças relacionadas à alimentação. Já com o incentivo de práticas mais sustentáveis e na gestão de resíduos com uma cadeia de suprimentos ética no setor de alimentos, também afetaria positivamente o objetivo de Consumo e Produção Responsáveis.

Em contrapartida, com a utilização dos dados de forma a ignorar o aspecto equitativo, pode ferir o objetivo de Redução das Desigualdades, aumentando a disparidade na cadeia alimentar, marginalizando grupos sociais vulneráveis, podendo inclusive afetar pequenos produtores.

## **10.5. Viés e Discriminação**

Viés e discriminação estão relacionados à possibilidade de algoritmos e sistemas tomarem decisões que prejudiquem determinados grupos, resultando em um tratamento desigual devido a distorções nos dados ou nos tratamentos realizados.

Hargittai (2020) ressalta que rastros comportamentais em redes sociais de indivíduos de classes sociais mais privilegiadas possuem maior probabilidade de estarem mais presentes nos conjuntos de dados em relação a indivíduos de comunidades mais periféricas, enviesando os dados comportamentais nessas amostras. Isso reforça que análises realizadas com essas bases acabam privilegiando tipos de indivíduos com base em sua classe social, ou seja, uma parte da população, e por isso é importante estar ciente sobre quais vozes estão representadas nos conjuntos de dados.

Para mitigar esses viéses e discriminação nos conjuntos de dados, uma prática comum que poderia ser implementada seria o balanceamento dos dados, visando equilibrar as amostras e garantir uma representação mais justa. No entanto, ao balancear excluindo ou criando dados novos pode prejudicar análises gerais, comprometendo a integridade e distorcendo as relações presentes nos dados originais. Assim, a implementação de filtros que permitem uma análise mais granular torna-se uma possibilidade viável, sem comprometer a integridade dos dados.

Filtros com granularidade permitem a seleção de dados com base em critérios específicos, possibilitando uma análise em cenários selecionados, mas isso não impede que o usuário do sistema, por vieses próprios, selecione recortes por regiões que intencionalmente excluem populações periféricas, com a finalidade de favorecer indivíduos mais privilegiados, ressaltando a importância de uma equipe diversa para a utilização do sistema implementado com Big Data.

## **10.6. Discussão**

A discussão sobre a utilização de Big Data na análise e tomada de decisões e o impacto na sociedade abrange diversas dimensões éticas e sociais, desde a privacidade e proteção dos dados, transparência na utilização e armazenamento de informações, até

viéses e discriminação que favorecem indivíduos enquanto prejudica outros. Isso ressalta a necessidade de implementar regulamentações que protegem os usuários e promovem um uso ético dos dados, com responsabilidade e visando não apenas o lucro das empresas, mas benefícios para a sociedade como um todo.

Entretanto, é importante ressaltar que o impacto de medidas regulatórias também podem impactar negativamente o desenvolvimento acadêmico a longo prazo. Schroeder e Cowls (2014) indicam que as implicações éticas para a privacidade e a segurança dos dados são diferentes em pesquisas acadêmicas, que visam um conhecimento generalizado, e pesquisas aplicadas, que buscam explorar os dados para fins comerciais e influenciar o comportamento político ou social dos indivíduos por exemplo. O limite do valor dos dados possui pesos diferentes, já que restrições na utilização dos dados impactam a demanda por conjuntos de dados em larga escala para pesquisas acadêmicas, utilizadas para melhor desenvolver o conhecimento científico-social a longo prazo, provocando uma redução nos ganhos das pesquisas acadêmicas, e pesquisas aplicadas não sofrem esse impacto para usos comerciais ou não acadêmicos, por serem reduzidos ou coletados pela própria empresa ou órgão governamental.

Driscoll e Walker (2014) ressaltam ainda que pesquisas acadêmicas realizadas com Big Data enfrentam dificuldades no processo de coleta, gerando dados não estruturados de baixa qualidade, resultando em perdas irreversíveis dos mesmos, impondo barreiras que limitam o progresso de tais pesquisas, onde empresas privadas conseguem recursos para acessar e manipular esses dados para benefício próprio.

# **11. Plano de Comunicação**

Para a redução de ruídos de comunicação, incertezas na equipe e promover a motivação e a colaboração entre os integrantes, um plano de comunicação se faz necessário para documentar estratégias e orientações sobre como as informações serão compartilhadas dentro de um projeto e entre as partes interessadas (Baptista, 2009).

Este plano inclui a identificação dos públicos-alvo, determinar os objetivos, definir das mensagens-chave para cada grupo, a escolha dos canais de comunicação mais apropriados, e a programação de quando e como as comunicações ocorrerão, além de contemplar a monitorização da eficácia da comunicação e a adaptação da estratégia conforme necessário (Amorim, 2018).

## **11.1. Objetivo**

Com a estruturação dos objetivos, é relevante ressaltar que estes devem ser alinhados com os objetivos e a visão das partes interessadas (Baptista, 2009).

O Projeto de Integração, Gerenciamento e Análise de Big Data tem como objetivo assegurar boas relações entre as partes interessadas, motivando os integrantes da equipe de desenvolvimento, alinhando stakeholders sobre o escopo do projeto e as etapas necessárias para seu desenvolvimento, bem como atividades a serem desenvolvidas, em desenvolvimento, finalizadas e atrasadas.

Isso é realizado por meio de uma comunicação transparente sobre o progresso, desafios, alterações e resultados do projeto, contribuindo assim para a possibilidade de colaboração entre os stakeholders e decisões embasadas no estado real do projeto.

## **11.2 Stakeholders**

Qualquer indivíduo ou grupo que pode influenciar ou ser influenciado pelos objetivos do projeto pode ser considerado um stakeholder, que, a partir de características e atributos próprios, possuem impactos diferentes no andamento do projeto, necessitando assim de estratégias diferentes para envolver as partes no seu desenvolvimento (Lyra et al, 2009).

Como o Projeto de Integração, Gerenciamento e Análise de Big Data possui duas organizações principais, por parte da instituição educacional de ensino superior Inteli e o parceiro de projeto a empresa de consultoria Integration Consulting, podemos separar os stakeholders em dois grupos:

### **11.2.1. Grupo Integration Consulting**

A equipe da Integration é liderada por Guilherme Paz, com André Durso como ponto focal backup, Jorge Jamil no comando técnico e Ian Matiussi na liderança de negócios e executiva, incluindo o onboarding executivo.

### **11.2.2. Grupo Inteli**

Dentro deste grupo temos os integrantes da equipe de desenvolvimento, que por serem alunos da instituição, possuem todos um peso de comunidade por desempenharem papéis semelhantes dentro da organização, bem como os professores instrutores do projeto, por fim o professor orientador do projeto, e o professor coordenador do curso de Sistemas de Informação.

### **11.2.3. Tipologia de stakeholders**

É possível avaliar os stakeholders a partir de três atributos, segundo Mitchell et al (1997):

- Poder: Possibilidade que um ator tem de realizar sua própria vontade a partir de forças simbólicas, provenientes de fontes diversas.
- Legitimidade: Ações são tomadas a partir de normas e definições por um sistema social.
- Urgência: Criticidade de tempo em relação às demandas.

De acordo com a combinação dos atributos destacados acima aos stakeholders, a classificação segue como na tabela a seguir, ressaltando que um stakeholder pode ser classificado por possuir um atributo que se sobressai aos demais, não significando que possui apenas um ou dois dos demais:

Tabela 1: Classificação de Stakeholders

Poder	Legitimidade	Urgência	Classificação	Definição
X	X	X	Definitivo	Stakeholder requer atenção e prioridade imediata pelo poder, possuir legitimidade para tal e a urgência das demandas
X	X	-	Dominante	Requer atenção pelo poder de influência e pela legitimidade das ações provenientes desse stakeholder
X	-	X	Perigoso	Por não possuir legitimidade, mas possuir poder e urgência, este stakeholder possui uma influência potencialmente coercitiva
-	X	X	Dependente	Apesar de possuir a legitimidade e a urgência das demandas, depende do poder de outro stakeholder para dar peso às suas reivindicações
X	-	-	Adormecido	Stakeholder que é necessário ser monitorado, enquanto não adquirir outro atributo, mas permanecendo com pouca ou nenhuma interação
-	X	-	Arbitrário	Não tem poder de influência, nem requisita urgência, requisitando atenção a respeito de sua posição na organização
-	-	X	Reivindicador	Deve ser monitorado caso adquira mais um atributo, mas

				não apresenta grande preocupação para a organização
--	--	--	--	---

Fonte: Adaptado adaptado de Mitchell et al. (1997).

#### 11.2.4. Classificação dos stakeholders

De acordo com a tabela 1, é possível classificar os stakeholders, separando-os nos dois grupos:

##### 11.2.4.1 Stakeholders Integration Consulting

Tabela 2: Categorização a partir do impacto no projeto dos stakeholders do grupo Integration Consulting.

Stakeholder	Poder	Legitimidade	Urgência	Classificação
Guilherme Paz/André Durso	X	X	X	Definitivo
Jorge Jamil		X		Arbitrário
Ian Matiussi		X		Arbitrário

Fonte: Autoria Própria a partir do modelo de Mitchell et al. (1997).

##### 11.2.4.2 Stakeholders Inteli

Tabela 3: Categorização a partir do impacto no projeto dos stakeholders do grupo Inteli.

Stakeholder	Poder	Legitimidade	Urgência	Classificação
Equipe de alunos		X	X	Dependente
Professores Instrutores		X		Arbitrário
Professor Orientador	X	X		Definitivo
Coordenador do Curso de SI	X	X	X	Definitivo

Fonte: Autoria Própria a partir do modelo de Mitchell et al. (1997).

## **11.3. Mensagens-chave**

Mensagens-chave devem apresentar temas relevantes para as partes interessadas no projeto, transmitindo a missão, posicionamento, objetivo e valores dos stakeholders. Se bem elaboradas, elas imprimem um sentido unificado, evitando os sentidos difusos (Lyra, 2021).

### **11.3.1. Mensagem-Chave para Integration Consulting**

“Somos pragmáticos e entendemos que a conexão com a rotina dos nossos clientes nos permite agir, com planejamento, e orientados ao resultado, integrando inovações tecnológicas às nossas soluções, inspirando movimento e preparando-os para o futuro.”

Palavras-chave: Pragmático, planejamento, resultado, inovação, tecnologia, resultado, inspiração, futuro.

### **11.3.2. Mensagem-Chave para Inteli**

“Somos movidos pela sede de conhecimento, nos adaptamos aos desafios, e entregamos à sociedade soluções com o potencial transformador, com o impacto que apenas os líderes do futuro podem oferecer, respeitando os princípios éticos, com integridade e responsabilidade.”

Palavras-chave: Conhecimento, adaptação, desafio, sociedade, impacto, liderança, futuro, ética, responsabilidade.

## **11.4. Canais de comunicação**

A escolha dos canais de comunicação se faz importante de acordo com a situação e o receptor da mensagem, considerando que foram utilizados majoritariamente Canais de Comunicação Pessoais (Oliveira, 2013).

Por padrão, a equipe de alunos que desenvolvem o projeto, bem como os professores instrutores, professor orientador e coordenador do curso utilizam a ferramenta Slack, para comunicar avanços no desenvolvimento do projeto, artefatos desenvolvidos,

links para estudo e pesquisa, comunicar obstáculos, e programar reuniões com um ou mais partes interessadas do projeto, limitando apenas ao grupo de stakeholders Inteli.

Também são realizadas as cerimônias da metodologia ágil Scrum, como Dailies, Plannings, e Retrospectivas, que envolvem o planejamento de Sprints, atribuição de atividades, comunicar impedimentos, e planos para superar desafios enfrentados entre os alunos da equipe de desenvolvimento e a própria turma do curso de Sistemas de Informação.

Com os parceiros de projeto, stakeholders da empresa Integration Consulting, os alunos da equipe desenvolvedora não possuem canal de comunicação aberto diretamente entre eles, necessitando contato através de professores instrutores ou o professor orientados. O contato com o parceiro de projeto acontece na cerimônia de Sprint Review, ao final de cada Sprint, com os incrementos do projeto, com coleta de feedbacks e apresentação do status report, bem como os próximos passos.

Esta rotina possibilita o desenvolvimento de maneira a validar o que será desenvolvido, e ajustar os incrementos até a Sprint final, com o foco em polir o que foi realizado com os feedbacks finais.

## 11.5. Plano de implementação

De acordo com o que deve ser desenvolvido no projeto, com os artefatos a serem feitos, e que vão ser incrementados, a relevância de elaborar um plano de comunicação para assegurar a resolução de conflitos e incentivar a troca de informação entre os integrantes da equipe e as outras partes interessadas. Para isso, o plano de implementação envolve três fases: Preparação, Execução e Monitoramento.

Na Preparação, as agendas são definidas para delimitar os prazos de entregas e são comunicadas as disponibilidades dos integrantes para com o desenvolvimento do projeto. Com isso, templates de comunicação são alinhados para criar meios de troca de mensagens, informações, definir responsabilidades e configurar ferramentas digitais necessárias para o desenvolvimento.

Já com Execução, iniciam-se as comunicações regulares e mantêm-se os registros, para viabilizar a coleta de dados para, posteriormente, utilizar nas métricas de desempenho dos integrantes. Nesta etapa também são feitos ajustes de acordo com necessidades pontuais dos integrantes, e o estabelecimento de pontos críticos que devem ser levados ao professor orientador ou instrutor.

Com o Monitoramento verificações regulares da eficácia e aderência ao plano de comunicação são realizadas, utilizando os dados gerados pelos integrantes, avaliando a eficácia do desenvolvimento e pontos de atenção para melhorias, elaborando feedbacks para os integrantes da equipe, visando o crescimento profissional e acadêmico, bem como avaliar o impacto das atividades na qualidade dos incrementos do projeto, alinhando com o professor orientador ajustes necessários no grupo para melhor performar no desenvolvimento do projeto.

## 11.6. Medidas de sucesso, Feedback e Ajustes

As medidas de sucesso são avaliadas pelo engajamento nas reuniões e comunicações, feedback dos stakeholders, cumprimento dos prazos estabelecidos e a satisfação geral dos stakeholders. Além da qualidade dos artefatos desenvolvidos, também é avaliada a rotação dos integrantes do grupo em atividades diferentes das que foram desenvolvidas por eles, visando uma integração maior entre cada integrante com diferentes etapas do projeto, ressaltando pontos de dificuldade enfrentados.

Uma medida avaliativa implementada na equipe é a Avaliação 360 graus, com a discussão sobre o desempenho de cada membro da equipe, com pontos com impacto positivo, pontos a melhorar, e o feedback de cada um, relacionando à qualidade de entrega e a participação em cada atividade, visando a rotação dos integrantes.

Nesse momento são utilizadas as evidências coletadas durante a fase de execução do plano de comunicação, discutindo de forma equilibrada diversos pontos da sprint, analisando a evolução de cada integrante.

Com os feedbacks, ajustes são realizados durante a Planning da Sprint seguinte, levando em conta a avaliação dos artefatos pelos professores de cada matéria, a validação das tecnologias e ideias de negócios com clientes, seguida por análises e ajustes periódicos nas estratégias de comunicação. A flexibilidade para alterar o plano de

comunicação conforme as necessidades do projeto e dos stakeholders é essencial para o sucesso contínuo, e também é um atributo da metodologia ágil.

## **12. Anexos**

Nesta seção apresenta-se o espaço destinado a informações complementares e relevantes ao conteúdo principal do projeto, utilizado para reforçar a argumentação do documento e contribuir para o entendimento completo.

### **12.1. Matriz de risco**

A matriz de risco é uma ferramenta para identificar e avaliar potenciais riscos que possam impactar negativamente no desenvolvimento do projeto. Neste tópico do anexo apresenta-se o histórico da matriz de risco utilizada em cada sprint do projeto, desde o seu início até o momento atual.

Cada sprint do projeto é acompanhada de uma matriz de risco específica, que é atualizada de acordo com as mudanças e imprevistos que surgem durante o planejamento do projeto. O objetivo desta seção de anexo é fornecer uma visão geral das matrizes de risco, permitindo uma análise comparativa do nível de risco enfrentado em cada momento.

## **13. Conclusões**

A engenharia de dados representa um avanço significativo na transformação dos processos de análise e uso de dados em organizações, e ao longo do desenvolvimento deste projeto, foi realizada a implementação de uma estrutura apta a lidar com a complexidade e o volume de dados, enquanto fornece informação para análises sobre o negócio.

O emprego da arquitetura de solução Big Data, não apenas permitiu a construção de um ambiente para manipulação e processamento de dados com a integração de diferentes fontes de dados e a automação de processos proporcionando uma visão mais detalhada do panorama da empresa.

Além disso, os ganhos em termos de segurança, escalabilidade e agilidade na análise e visualização de informações não podem ser subestimados, visto que a qualidade dos dados, pode impulsionar tomadas de decisões mais assertivas e fomentar a inovação em nas estratégias de marketing, vendas e, consequentemente, no crescimento do negócio.

E o futuro onde a informação é um pilar essencial para a excelência operacional e o sucesso empresarial, este projeto promove o diferencial competitivo.

## **14. Referências**

AMORIM, M. Plano de Comunicação: APICCAPS, 2018. Trabalho de Mestrado em Marketing - Universidade Católica Portuguesa.

BAPTISTA, A. Plano De Comunicação Interna Para A Sonae Sierra, 2009. Projeto de Mestrado em Gestão de Empresas - Instituto Superior de Ciências do Trabalho e da Empresa, Lisboa.

DALFOVO, M. NUNCIO, C. Plano de comunicação para posicionamento da marca Bella Janela Indústria de Cortinas Ltda, 2009. Revista Interdisciplinar Científica Aplicada, Blumenau.

D'AMARIO, E.; SORANZ, R. A aplicação do modelo de saliência de stakeholders em gestores de bancos de varejo, 2014. Encontro Internacional Sobre Gestão Empresarial E Meio Ambiente.

Driscoll, K., & Walker, S. (2014). "Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data." International Journal of Communication.

Hargittai, E. (2020). "Potential Biases in Big Data: Omitted Voices on Social Media." Social Science Computer Review.

LI, J., YANG, J., HERTZMANN, A., ZHANG, J., XU, T. Layoutgan: Generating Graphic Layouts With Wireframe Discriminators. 2019 - Beijing Institute of Technology. Disponível em: <https://arxiv.org/pdf/1901.06767.pdf>

LYRA, J. Comunicação estratégica para transformar: a importância da mensagem-chave para um público segmentado no contexto das ONG e o caso da Mundo A Sorrir, 2021. Mestrado em Ciências da Comunicação, Especialização em Publicidade e Relações Públicas, Universidade do Minho.

LYRA, M. et al. O Papel dos Stakeholders na Sustentabilidade da Empresa: Contribuições para Construção de um Modelo de Análise, 2009. RAC, Curitiba.

MITCHELL, R. et al. Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts, 1997. Academy of Management Review.

OLIVEIRA, P. Plano de Comunicação Empresarial para a empresa Ultramaisfarma, 2013. Centro de Ensino Superior do Ceará, Fortaleza.

Possamai, A. J. (2016). "Dados Abertos no Governo Federal Brasileiro: Desafios de Transparência e Interoperabilidade." Dissertação (Mestrado em Ciência Política), Instituto de Filosofia e Ciências Humanas, Programa de Pós-Graduação em Ciência Política, Universidade Federal do Rio Grande do Sul.

Schroeder, R., & Cowls, J. (2014). "Big Data, Ethics, and the Social Implications of Knowledge Production." Oxford Internet Institute.

Tene, O., & Polonetsky, J. (2013). "Big Data for All: Privacy and User Control in the Age of Analytics." Northwestern Journal of Technology and Intellectual Property.

Taylor, L. (2017). "What is data justice? The case for connecting digital rights and freedoms globally." Big Data & Society.

Varley-Winter, O., & Shah, H. (2016, December 28). "The Opportunities and Ethics of Big Data: Practical Priorities for a National Council of Data Ethics.", Royal Statistical Society.

White, G., Ariyachandra, T., & White, D. (2019). "Big Data, Ethics, and Social Impact Theory – A Conceptual Framework." The Journal of Management and Engineering Integration, 12(1), 9. Xavier University.

ZHANG, M. Speeding Up The Prototyping Of Low-Fidelity User Interface Wireframes. 2022 - Research Compliance Certification. Disponível em: <https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/196603/ZHANG-FINALTHESIS-2022.pdf?sequence=1&isAllowed=y>

## Links usados:

AMAZON WEB SERVICES, INC. What is Amazon Redshift?. Disponível em: <https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html>.

STACK OVERFLOW. Columnar database queries in Amazon Redshift. Disponível em: <https://stackoverflow.com/questions/45176680/columnar-database-queries-in-amazon-redshift>.

AMAZON WEB SERVICES, INC. Columnar storage. In: *Amazon Redshift*. Disponível em: [https://docs.aws.amazon.com/redshift/latest/dg/c\\_columnar\\_storage\\_disk\\_mem\\_mgmt.html](https://docs.aws.amazon.com/redshift/latest/dg/c_columnar_storage_disk_mem_mgmt.html).

AMAZON WEB SERVICES, INC. Amazon Redshift deep dive - Data Warehousing on AWS. Disponível em: <https://docs.aws.amazon.com/redshift/latest/dg/welcome.html>.

AMAZON WEB SERVICES, INC. Amazon Redshift - Big Data Analytics Options on AWS. Disponível em: <https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html>.

TO THE NEW BLOG. Amazon Redshift: A Comprehensive Overview. Disponível em: <https://www.tothenew.com/blog/amazon-redshift-a-comprehensive-overview/>.

Fontes relacionadas ao desenvolvimento do TAM:

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

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

Fontes relacionadas ao desenvolvimento do SAM:

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

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

<https://www.cortex-intelligence.com/intelligence-review/varejo-%C3%A9-o-segmento-que-mais-investe-em-tecnologia-no-brasil>

Fontes relacionadas ao desenvolvimento do SOM:

<https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca%20de%2028%2C6%2520da,e%20Estat%C3%ADstica%20>