



BIG DATA

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA

INTEGRATION

Autores:

Gabriel Santos

Henri Harari

Rafael Moritz

Patrick Miranda

Raduan Muarrek

Vitor Moura

Data de criação: 24 de Outubro de 2023

SÃO PAULO – SP

2023

Controle de Documento

Histórico de Revisões

Data	Autor	Versão	Resumo da atividade
26/1 0/2023	Patrick	1.0	Adição dos artefatos da sprint 1: Artefatos de análise de negócio, Análise de Experiência do usuário e arquitetura da solução.
09/11 /2023	Henri	1.1	Adição da documentação do wireframe.
12/11 /2023	Patrick	2.0	Revisão da formatação do arquivo para entrega
13/11 /2023	Vitor	2.0	Revisão da formatação do arquivo para entrega
24/11 /2023	Patrick	3.0	Adição de análise ética do projeto Adição de documentação sobre data warehouse(Redshift) do projeto

Table 1: Controle de documento

Sumário

Controle de Documento	3
Sumário	3

1. Introdução	5
1.1 Parceiro de Negócios	6
1.2 Problema	6
1.2.1 Definição do Problema	6
2. Objetivos	7
2.1 Objetivos Gerais	7
2.2 Objetivos Específicos	7
2.3 Justificativa	7
3. Análise de Negócios	8
3.1 Proposta de Valor	8
3.2 Matriz de Risco	9
3.3 TAM SAM SOM	10
4. Análise de Experiência do Usuário	14
4.1 Personas	14
4.2 Jornada do Usuário	15
4.3 User Stories	16
5. Wireframe	21
5.1 Telas	21
5.1.1 Dashboard	21
5.1.2 Infográfico	22
8.2 Requisitos a cumprir	23
5.3 Justificativa das Escolhas de Design	24
8.5 Feedback e Iterações	28
5. Análise Exploratória e fonte dos dados	29
5.1 Introdução	29
5.2 Método	32
6. Arquitetura Macro	33
6.1 Identificação dos dados:	33
6.1.1 Dados públicos:	33
6.1.2 Dados do cliente:	35
6.2 Gestão de dados:	35
6.2.1 Dados públicos:	35
6.2.2 Dados privados:	36
6.3 Seleção dos serviços AWS:	36
6.4 Fluxo dos dados:	37
6.5 Segurança:	38
6.6 Monitoramento e gerenciamento:	38
7. Data lake	39
7.1 Introdução	39
7.2 S3	39
7.3 Buckets	39

7.3.1 Configuração dos Buckets	40
9. Conclusões	43
10. Referências	44
11. Anexos	45

1. Introdução

1.1 Parceiro de Negócios

1.2 Problema

1.2.1 Definição do Problema

2. Objetivos

2.1 Objetivos Gerais

2.2 Objetivos Específicos

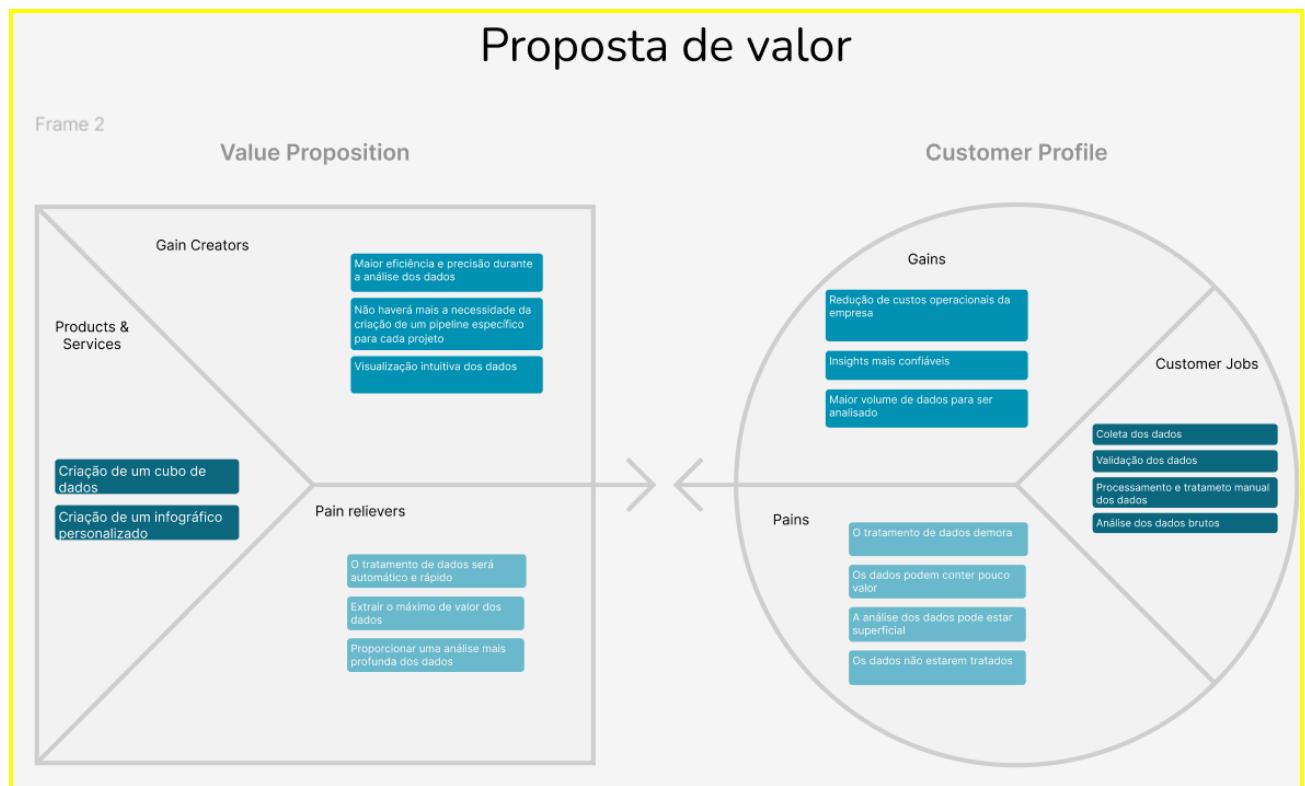
2.3 Justificativa

3. Análise de Negócios

3.1 Proposta de Valor

Ao usar o Value Proposition Canvas, as empresas podem alinhar suas ofertas com as necessidades do cliente, melhorar suas mensagens de marketing e se diferenciar de seus concorrentes.

imagem 1: Canvas proposta de valor



Fonte: Dados dos autores (2023)

Pains (Dores) e Pain Relievers (Aliviadores de Dor):

As "Dores" representam os desafios específicos que consultorias enfrentam ao tentar entender o mercado na indústria alimentícia, como a demora na coleta e análise de dados, a falta de insights relevantes e a dificuldade em interpretar grandes volumes de dados brutos. Os "Aliviadores de Dor" demonstram como o pipeline de Big Data pode solucionar esses desafios. Isso inclui um tratamento de dados mais ágil, extração de

insights valiosos de vastos conjuntos de dados e a capacidade de realizar análises mais profundas e precisas.

Gains (Ganhos) e Gain Creators (Criadores de Ganho):

Os "Ganhos" expressam os resultados positivos e benefícios que as consultorias desejam alcançar, como otimização de estratégias de "go to market", maior eficiência operacional e geração de insights inovadores para seus clientes na indústria alimentícia. Os "Criadores de Ganho" elucidam como o pipeline de Big Data facilita esses ganhos. Isso pode incluir uma análise mais precisa dos hábitos do consumidor, tendências emergentes no mercado alimentício e identificação de oportunidades inexploradas.

Products & Services (Produtos e Serviços) e Customer Jobs (Trabalhos do Cliente):

Conclusão

A seção "Produtos e Serviços" destaca as soluções específicas proporcionadas pelo pipeline, como ferramentas de visualização de dados, análise preditiva e segmentação avançada do mercado. Os "Trabalhos do Cliente" representam as tarefas ou atividades que as consultorias precisam realizar, como entender padrões de consumo, identificar novos nichos de mercado e formular estratégias de penetração de mercado eficazes. Em síntese, o canva "Proposta de Valor" para este projeto de pipeline de Big Data busca direcionar e articular o valor tangível oferecido às consultorias voltadas para a indústria alimentícia. Ele ilustra como a solução aborda dores específicas do mercado, potencializa

3.2 Matriz de Risco

É uma das principais ferramentas na análise de negócios, utilizada para o gerenciamento de riscos operacionais existentes na empresa. A Figura 2, ilustra a construção da matriz de risco para o projeto.

imagem 1: matriz de risco

Matriz de Risco											
Probabilidade	Riscos						Oportunidades				
	Muito Alta	1							Reconhecimento no mercado pela solução inovadora		
	Alta	2			Erros na análise estatística	Falha na integração entre serviços			Integração mais profunda com outras ferramentas	Otimização dos custos	Melhoria contínua através do feedback dos usuários finais
	Média	3			Limitações com ferramentas open-source	Falha na ingestão de dados	Custo inesperado		Descoberta de novas fontes de dados relevantes		
	Baixa	4			Problemas na visualização.	Dados incompletos ou corrompidos	Falhas no processamento	Exceder o tempo estipulado para entrega do MVP		Expansão da solução para outros distribuidores e mercados	Adoção de novas ferramentas open-source
	Muito Baixa	5						Falhas de segurança nos dados			
		1	2	3	4	5	5	4	3	2	1
		Muito Baixa	Baixa	Média	Alta	Muito Alta	Muito Alta	Alta	Média	Baixa	Muito Baixa
		Impacto									

Fonte: Dados dos autores (2023)

Enquanto a matriz de risco proporciona uma visão clara das possíveis contingências e desafios que podem surgir, abaixo apresenta-se o plano de ação onde representa a resposta estratégica para enfrentar tais eventualidades.

imagem 2: plano de ação

Plano de Ação						
Matriz	Descrição do Impacto	Responsável	Ação	Descrição da Ação	Previsão	
2-3	Erros na análise estatística	Rafael	Mitigar	Revisar métodos e validar com especialistas.	Sprint 4	
2-4	Falha na integração entre serviços.	Patrick	Mitigar	Testes de integração frequentes.	Sprint 3	
3-3	Limitações com ferramentas open-source.	Patrick	Prevenir	Pesquisa prévia e testes de adequação.	Sprint 3	
3-4	Falha na ingestão de dados.	Henri	Mitigar	Monitorar processos de ingestão e ter backups.	Sprint 2	
3-5	Custo inesperado.	Raduan	Mitigar	Monitorar e controlar gastos.	Sprint 3	
4-3	Problemas na visualização.	Vitor	Mitigar	Treinamento na ferramenta e testes frequentes.	Sprint 5	
4-4	Dados incompletos ou corrompidos.	Gabriel	Prevenir	Validação dos dados antes da ingestão.	Sprint 2	
4-4	Falhas no processamento.	Henri	Mitigar	Monitorar a performance e otimizá-la.	Sprint 4	
4-5	Exceder o tempo estipulado para entrega do MVP.	Gabriel	Prevenir	Gestão eficaz do projeto e revisões frequentes.	Sprint 5	
5-5	Falhas de segurança nos dados.	Raduan	Prevenir	Implementar protocolos de segurança rigorosos.	Sprint 3	

Fonte: Dados dos autores (2023)

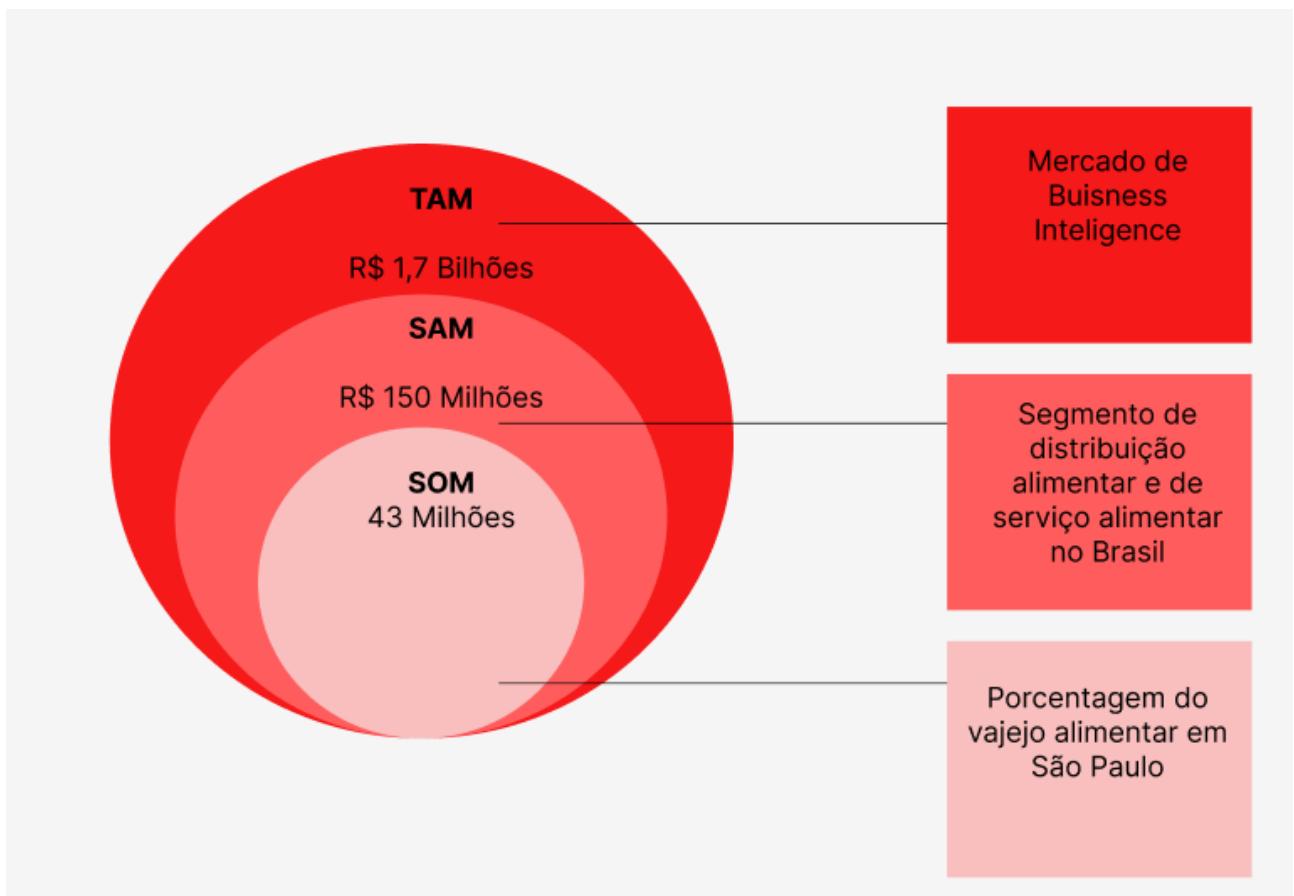
3.3 TAM SAM SOM

TAM (Total Addressable Market): Representa o mercado total que poderia se beneficiar ou necessitar do seu serviço ou produto. É o valor total caso 100% do mercado adotasse seu produto.

SAM (Serviceable Addressable Market): É a parcela do TAM que realmente pode ser alcançada por seu produto ou serviço, levando em consideração as limitações geográficas, de distribuição, capacidade e outras.

SOM (Serviceable Obtainable Market): Representa a porção do SAM que se espera atingir em um determinado período de tempo, considerando fatores como concorrência, barreiras de entrada e estratégia de implementação.

Imagen 1: TAM SAM SOM



TAM

Descrição: Total estimado de receita no mercado de Business Intelligence relacionado à indústria alimentícia.

Premissa: Existe uma grande demanda por insights e análises na indústria alimentícia. O TAM engloba todas as consultorias, empresas e indivíduos que poderiam

potencialmente se beneficiar do uso de ferramentas de Business Intelligence na indústria alimentícia.

Valor: R\$ 1,7 bilhões.

SAM

Descrição: Total estimado de receita que pode ser direcionada pelas consultorias que servem o segmento de distribuição alimentar e serviço alimentar no Brasil utilizando insights de big data.

Premissa: Dentro do amplo mercado de Business Intelligence para a indústria alimentícia, há um segmento específico focado na distribuição e no serviço alimentar que pode ser diretamente beneficiado por insights de big data. Este segmento tem necessidades mais específicas e pode ser atendido de forma mais direcionada pelo seu serviço.

Valor: R\$ 150 milhões.

SOM

Descrição: Receita potencial estimada que pode ser capturada pelas consultorias em um determinado período de tempo, focando especificamente no varejo alimentar de São Paulo.

Premissa: São Paulo, sendo um grande hub comercial, possui um segmento significativo de varejo alimentar. Focar neste segmento proporciona uma oportunidade tangível e mensurável. A premissa é que, ao focar em um mercado específico e conhecido, como o varejo alimentar de São Paulo, você pode oferecer soluções mais personalizadas e alcançar uma maior penetração de mercado.

Valor: R\$ 43 milhões.

4. Análise de Experiência do Usuário

4.1 Personas

A persona é uma representação humanizada do público-alvo ideal e é usada para ajudar a equipe de desenvolvimento a compreender melhor suas necessidades, desejos e comportamentos. No projeto atual, foram identificadas duas personas, o tech lead, responsável pela construção do cubo de dados e o consultor de marketing, responsável pela análise do cubo.

Imagen 1: Persona, Moisés Aragão



Fonte: Dados dos autores (2023)

Imagen 2: Persona, Enzo Ananias



Enzo Ananias

Consultor de Marketing - Integration

- 25 anos
- Análise de tendências
- Inovador
- Beach tennis
- Solteiro
- "Não há tempo para perder"



01

KPIs.

- Lucro real da empresa consultada
- Qualidade dos insights
- Eficiência das análises

02

Desejos

- Eficiência Operacional
- Acompanhamento de Tendências

03

Necessidades:

- Praticidade para analisar os dados
- Insights de melhor qualidade

04

Dores:

- Grande volume de dados
- Repetição de tarefas analíticas

05

Objetivos (Momento que usará o sistema):

- Analisar grande volume de dados
- Fornecer insights confiáveis

Fonte: Dados dos autores (2023)

4.2 Jornada do Usuário

A jornada do usuário é uma representação visual ou narrativa do percurso que um indivíduo realiza ao interagir com um produto, serviço ou sistema, desde o primeiro contato até a conclusão de um objetivo específico, levando em consideração suas emoções, experiências e desafios ao longo do caminho. Ela ajuda a compreender as necessidades, motivações e pontos de atrito do usuário, facilitando a criação de experiências mais eficientes e satisfatórias.

Imagen 1: jornada, Enzo Ananias



Fonte: Dados dos autores (2023)

Imagen 2: jornada, Moisés Aragão



Fonte: Dados dos autores (2023)

4.3 User Stories

Abaixo seguem quatro user stories realizados no padrão INVEST, para garantia do padrão de qualidade. Duas referentes ao teach lead e duas referentes ao consultor de marketing.

Número	US01
Título	Atualizar dados Públicos do IBGE
Pessoas	Moisés Aragão
História	Eu como teach lead, quero atualizar os dados do IBGE, de forma a garantir que os dados utilizados para análise estejam condizentes com o último censo.
Critérios de aceitação	<ol style="list-style-type: none"> 1. Validar formato da planilha .csv 2. Validar valores da planilha .csv 3. Efetuar atualização dos dados em .csv para o S3.
Testes de aceitação	<p>Critério 1:</p> <ul style="list-style-type: none"> - A planilha está no formato .csv <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério. - Recusou: Errado, analisar o processamento da planilha enviada. - O nome da planilha corresponde à lista de planilhas esperadas pelo pipeline <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério. - Recusou: Errado, analisar nome e fonte da planilha enviada. <p>Critério 2:</p> <ul style="list-style-type: none"> - Dados duplicados ou corrompidos foram removidos. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, tratar os dados <p>Critério 3:</p> <ul style="list-style-type: none"> - Os novos dados governamentais foram enviados para atualização do S3. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, analisar formato e origem dos dados - As colunas das planilhas no S3 não estão traduzidas: <ul style="list-style-type: none"> - Aceitou: Errado, revisar tratamento antes do consumo - Recusou: Correto, dados atualizados com sucesso.

Nú mero	US02
Títu lo	Atualizar dados Públicos de CNPJ
Per sonas	Moisés Aragão
Hist ória	Eu como teach lead, quero atualizar os dados de CNPJ, de forma a garantir que os dados utilizados para análise estejam condizentes com o último censo.
Crit érios de aceitação	<ol style="list-style-type: none"> 1. Validar formato da planilha .csv 2. Validar valores da planilha .csv 3. Efetuar atualização dos dados em .csv para o S3.
Test es de aceitação	<p>Critério 1:</p> <ul style="list-style-type: none"> - A planilha está no formato .csv <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério. - Recusou: Errado, analisar o processamento da planilha enviada. - O nome da planilha corresponde à lista de planilhas esperadas pelo pipeline <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério. - Recusou: Errado, analisar nome e fonte da planilha enviada. <p>Critério 2:</p> <ul style="list-style-type: none"> - Dados duplicados ou corrompidos foram removidos. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, tratar os dados <p>Critério 3:</p> <ul style="list-style-type: none"> - Os novos dados governamentais foram enviados para atualização do S3. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, analisar formato e origem dos dados - As colunas das planilhas no S3 não estão traduzidas: <ul style="list-style-type: none"> - Aceitou: Errado, revisar tratamento antes do consumo - Recusou: Correto, dados atualizados com sucesso.

Nú mero	US03
Títu lo	Disponibilizar visualização do infográfico
Per sonas	Enzo Ananias
Hist ória	<p>Eu, como consultor de marketing, quero visualizar o infográfico gerado com base no cubo de dados, para que eu possa ter auxílio na tomada de decisões .</p>
Crit érios de aceitação	<p>1. O usuário deve ser capaz de acessar a página do infográfico.</p> <p>2. O usuário deve ser capaz de exportar o infográfico.</p>
Test es de aceitação	<p>Critério 1:</p> <ul style="list-style-type: none"> - Usuário iniciou o acesso a página do infográfico. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, analisar origem do erro e resolvê-lo <p>Critério 2:</p> <ul style="list-style-type: none"> - Usuário efetuou a exportação do infográfico <ul style="list-style-type: none"> - Aceitou: Correto, infográfico exportado com sucesso - Recusou: Errado, analisar origem do erro e resolvê-lo

Nú mero	US04
Títu lo	Atualizar dados do cliente
Per sonas	Moisés Aragão
Hist ória	<p>Eu como teach lead, quero que os dados de vendas do cliente sejam semanalmente atualizados, garantindo que os dados analisados do cliente estejam sempre atualizados.</p>
Crit érios de aceitação	<ol style="list-style-type: none"> 1. O usuário consegue visualizar os dados do cliente . 2. Verificar se os dados da API do parceiro estão atualizados 3. Verificar se o consumo ocorreu automaticamente no horário definido.

Testes de aceitação	<p>Critério 1:</p> <ul style="list-style-type: none"> - A requisição foi aceita pela API do parceiro. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério. - Recusou: Errado, entrar em contato com o parceiro para discutir a disponibilidade da api. <p>Critério 2:</p> <ul style="list-style-type: none"> - Os dados presentes no cubo correspondem com o registro mais recente. <ul style="list-style-type: none"> - Aceitou: Correto, começar o próximo critério - Recusou: Errado, entrar em contato com o parceiro para solicitar revisão dos dados fornecidos. <p>Critério 3:</p> <ul style="list-style-type: none"> - A atualização semanal foi acionada automaticamente. <ul style="list-style-type: none"> - Aceitou: Correto, dados foram atualizados com sucesso. - Recusou: Revisar código de automatização
----------------------------	--

Número	US5
Título	Atualizar dados consumidos pelo infográfico.
Personas	Enzo Ananias
História	Eu como consultor de marketing, quero que o infográfico reflita a realidade do cubo de dados, de forma a garantir que os dados visualizados correspondam aos últimos dados do cliente.
Critérios de aceitação	<ol style="list-style-type: none"> 1. O usuário deve ter acesso aos dados pelo infográfico de forma correspondente ao cubo. 2. Os dados consumidos devem ser os mais recentes gerados.

Testes de aceitação	<p>Critério 1:</p> <ul style="list-style-type: none"> - Verificar se as informações no cubo e nos infográficos correspondem. <ul style="list-style-type: none"> - consistentes: Correto, começar o próximo critério - inconsistentes: Errado, revisar código de consumo dos dados para geração dos infográficos <p>Critério 2:</p> <ul style="list-style-type: none"> - Verificar se os infográficos foram devidamente atualizados junto ao cubo de dados. <ul style="list-style-type: none"> - Dados recentes: Correto, cubo de dados atualizados - Dados anteriores: Errado, revisar código de consumo dos dados para geração dos infográficos
----------------------------	--

5. Wireframe

Um wireframe de baixa fidelidade é uma representação esquemática e simplificada de uma interface ou design de um produto digital. Ele é geralmente criado no estágio inicial do processo de design para esboçar a estrutura básica e a disposição dos elementos, sem detalhes gráficos ou estilísticos. O foco está na funcionalidade e na organização da informação, permitindo uma avaliação rápida e fácil das ideias e conceitos do design. O wireframe da sprint 1 serviu de base para a validação com o parceiro de negócios e primeiro passo para suas posteriores iterações.

5.1 Telas

O objetivo estimado com as interfaces do wireframe é uma visualização fidedigna dos dados provenientes do pipeline de Big Data, contribuindo na tomada de decisão dos consultores, assim como o fácil entendimento das métricas dispostas. Para que esse objetivo seja alcançado dividimos a nossa aplicação em duas telas, o dashboard e o infográfico.

5.1.1 Dashboard

O Dashboard abriga a visualização inicial, onde temos um sidebar para a exposição de projetos anteriores, assim o consultor tem o fácil acesso como um hub a insights, antes gerados, que poderão ser incorporados ao próximo atendimento de consultoria. Posteriormente a seleção do projeto, abriga-se a tela principal, que apresentará pelo menos 8 gráficos distintos, cada um destacando diferentes informações do mercado ou do cliente. Segmentamos a tela estrategicamente a fim de uma melhor experiência do usuário, contendo áreas destinadas somente a dados do cliente, outra a dados do mercado e região e por último a pesquisa por setor. Finalmente, a Sidebar disponibilizará um botão de acesso ao Infográfico.

A partir do gráfico 1 esperamos, uma linha temporal de vendas para a visualização de alguma tendência de longo prazo, no gráfico 2 à 4 haverão métricas de número de CNPJs, número de clientes, número de vendas realizadas e potencial do consumidor para cada canal, categoria e região. A região do gráfico 5, seguindo a proposta do Wireframe, deverá ainda ser alinhada com o cliente. No gráfico 6, por sua vez, haverá a visualização do mapa de CNPJs do mercado Brasileiro de acordo com os CNAEs desejado para análise.

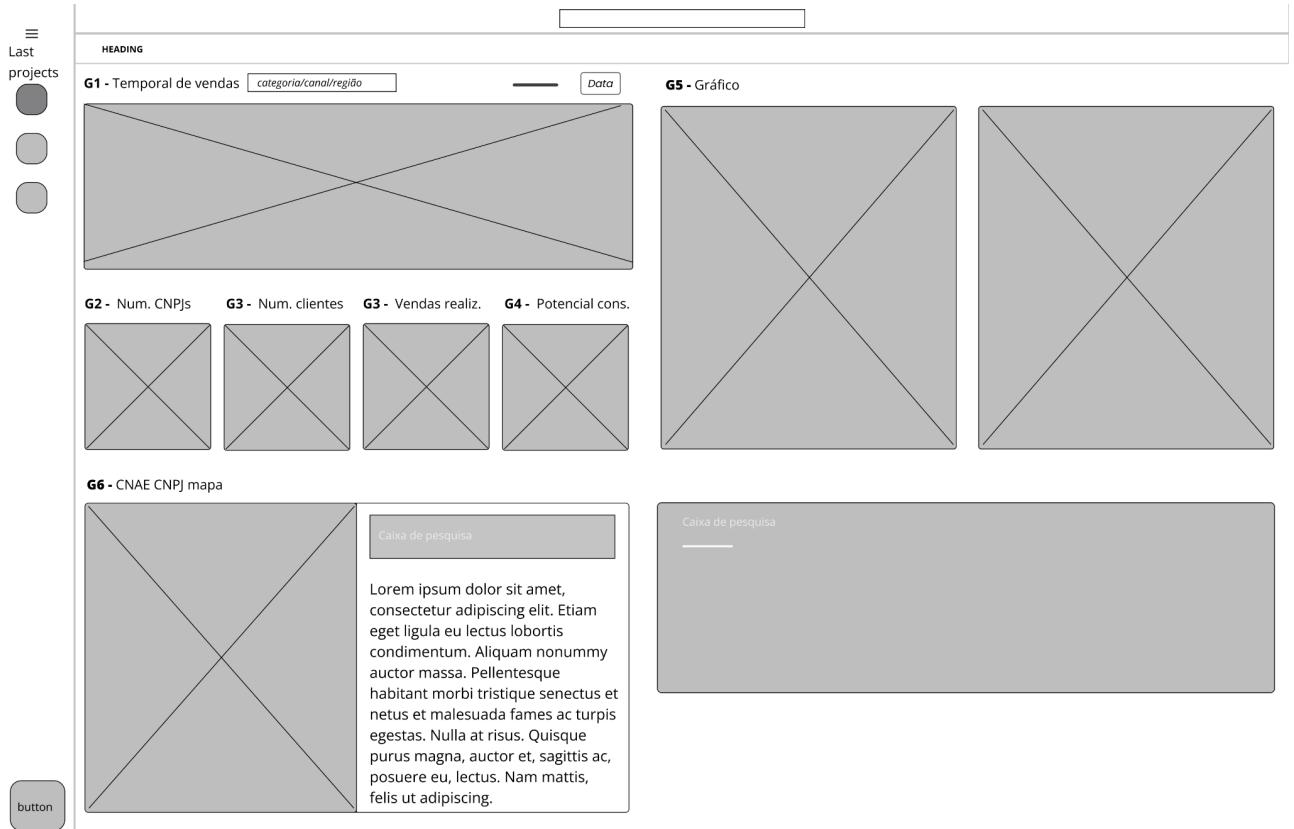


Figura XX - Dashboard Wireframe

Fonte: Elaborado pelo próprio autor (2023)

5.1.2 Infográfico

A tela de infográfico abriga a mesma Sidebar presente no dashboard, estando na tela principal as informações que ajudarão o consultor no entendimento das métricas e tendências. Para cada gráfico, presente na tela do dashboard, são descritas suas métricas e entregas de valor para a análise.

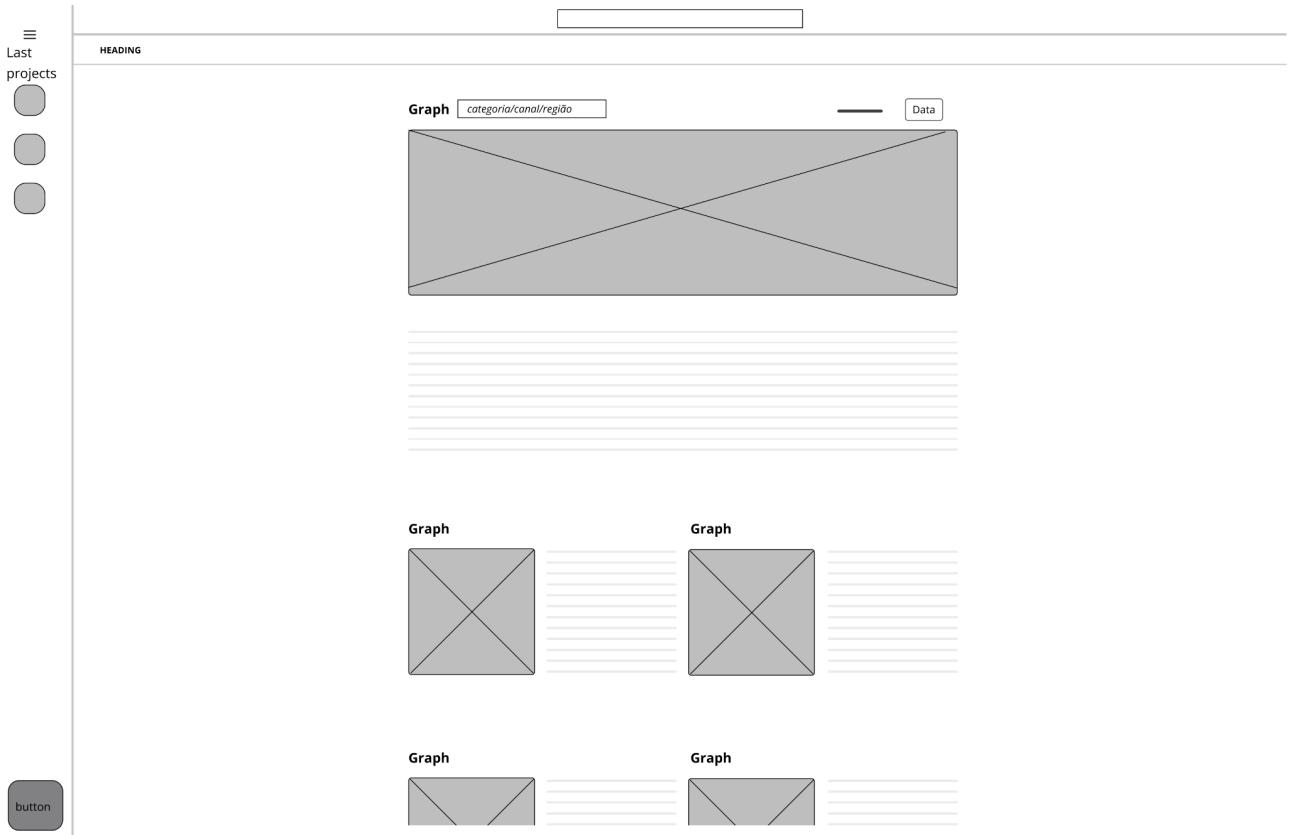


Figura XX - Infográfico Wireframe

Fonte: Elaborado pelo próprio autor (2023)

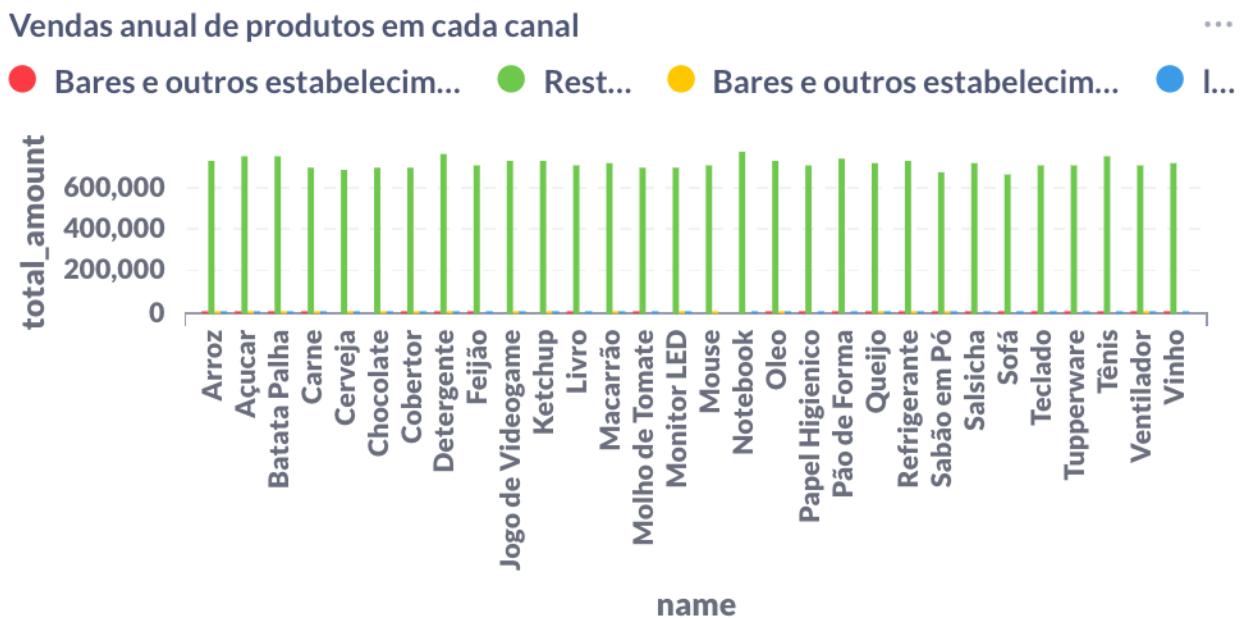
5.1.2.1 Gráfico de Mapa

O gráfico de mapa é uma ferramenta visual valiosa que representa dados geoespaciais de maneira intuitiva e compreensível. Ele permite a análise e visualização de informações em contextos geográficos, destacando padrões e correlações em regiões específicas. Sua utilidade abrange diversas aplicações, desde a representação de dados demográficos e distribuição de recursos até o monitoramento de atividades geográficas em tempo real. Ao apresentar dados de forma geográfica, o gráfico de mapa facilita a identificação de tendências espaciais, apoia a tomada de decisões estratégicas e proporciona uma compreensão visualmente impactante das complexidades geográficas inerentes aos conjuntos de dados.

Ao aplicar essa ferramenta, é possível representar de forma clara o número de CNPJs por canal/região, permitindo uma compreensão instantânea da distribuição geográfica dessas entidades. Além disso, o gráfico de mapa é útil para analisar o consumo por categoria, canal e região, destacando padrões de comportamento e preferências em diferentes áreas geográficas. Para entender as vendas por CNPJ em relação a cada categoria e região, o gráfico de mapa proporciona uma visão visualmente impactante da distribuição geográfica das transações comerciais. A análise DE-PARA Canal-CNAE, que associa canais a códigos CNAE, também pode ser eficientemente representada em um gráfico de mapa, possibilitando uma interpretação espacial das relações entre diferentes categorias de atividades econômicas e canais de distribuição. Essa abordagem visual enriquece a compreensão dos dados, facilitando a identificação de padrões e a tomada de decisões estratégicas com base em informações geográficas.

5.1.2.2 Gráfico de barras

Um gráfico de barras é uma representação visual que utiliza barras retangulares para exibir dados em categorias específicas, sendo especialmente eficaz para ilustrar a distribuição ou a comparação de quantidades. Cada barra no gráfico representa uma categoria e a altura da barra corresponde ao número total associado a essa categoria. Esse tipo de gráfico proporciona uma visão clara e comparativa das relações entre diferentes categorias e suas respectivas quantidades, tornando-o uma ferramenta valiosa para a análise visual de dados e a identificação de padrões ou tendências em conjuntos de informações categóricas.



5.1.2.2 Gráfico de linhas

Um gráfico de linhas é uma representação visual que exibe a relação entre duas variáveis ao longo de um eixo horizontal e vertical. No contexto de determinar o número de vendas de cada produto, um gráfico de linhas representaria a evolução ou variação das vendas ao longo do tempo ou de alguma outra dimensão relevante. Cada linha no gráfico representaria um produto específico, e os pontos ao longo da linha indicariam a quantidade de vendas para esse produto em momentos específicos. Esse tipo de gráfico é eficaz para identificar padrões de vendas, tendências sazonais e comparar o desempenho relativo de diferentes produtos ao longo do período analisado. A visualização clara e contínua das linhas facilita a interpretação dos dados de vendas e fornece insights valiosos para a análise de comportamento de compra de cada produto.

venda de categoria por mes top 7



5.2 Requisitos a cumprir

Para atender aos requisitos do cliente no desenvolvimento do cubo de dados e sua visualização no dashboard, o principal requisito é obter e integrar diversas informações essenciais do mercado e do cliente. Primeiramente, seguindo as expectativas apontadas pelo parceiro no workshop do dia 20/10/23, é necessário ter acesso ao número de CNPJs por canal e região, assim como os dados de consumo por categoria, canal e região. Além disso, a fim de permitir uma visualização regional no gráfico 6, será necessário o mapeamento entre os Canais e os códigos CNAE. Com base nesses dados, o objetivo final é criar um cubo que permita analisar o potencial de mercado de forma mais automatizada, desdobrando para cada combinação de categoria, canal e região: o número total de CNPJs, o número de clientes atendidos, as vendas realizadas e o potencial de consumo dispostos no primeiro quadrante do wireframe.

É importante ressaltar que, ao final do projeto, a entrega não consistirá em uma solução pronta, como um modelo preditivo, mas sim na estruturação e engenharia de dados necessárias para viabilizar uma solução que será implementada pela empresa no futuro. Portanto, o foco está na criação de uma infraestrutura sólida e eficiente para suportar a análise de mercado de forma automatizada e precisa.

5.3 Justificativa das Escolhas de Design

A justificativa das escolhas de design para essa tela é fundamentada na necessidade de proporcionar ao usuário uma experiência intuitiva e eficiente na visualização e interação com os dados apresentados. A presença de três gráficos - representando respectivamente a evolução temporal de vendas por canal, a distribuição geográfica dos canais e o relacionamento entre CNAE e CNPJ - foi determinada com base na relevância e na importância estratégica dessas informações para o usuário, de acordo com os seus desejos expressos no workshop.

Disponibilizando uma caixa de pesquisa para setores no canto inferior direito, será facilitada a navegação e a localização específica de informações para cada case de consultoria. Essa funcionalidade adiciona um componente de personalização e agilidade à experiência do usuário, permitindo-lhe encontrar dados pertinentes ao setor de atuação - Go-to-Market - do cliente específico de maneira mais granular.

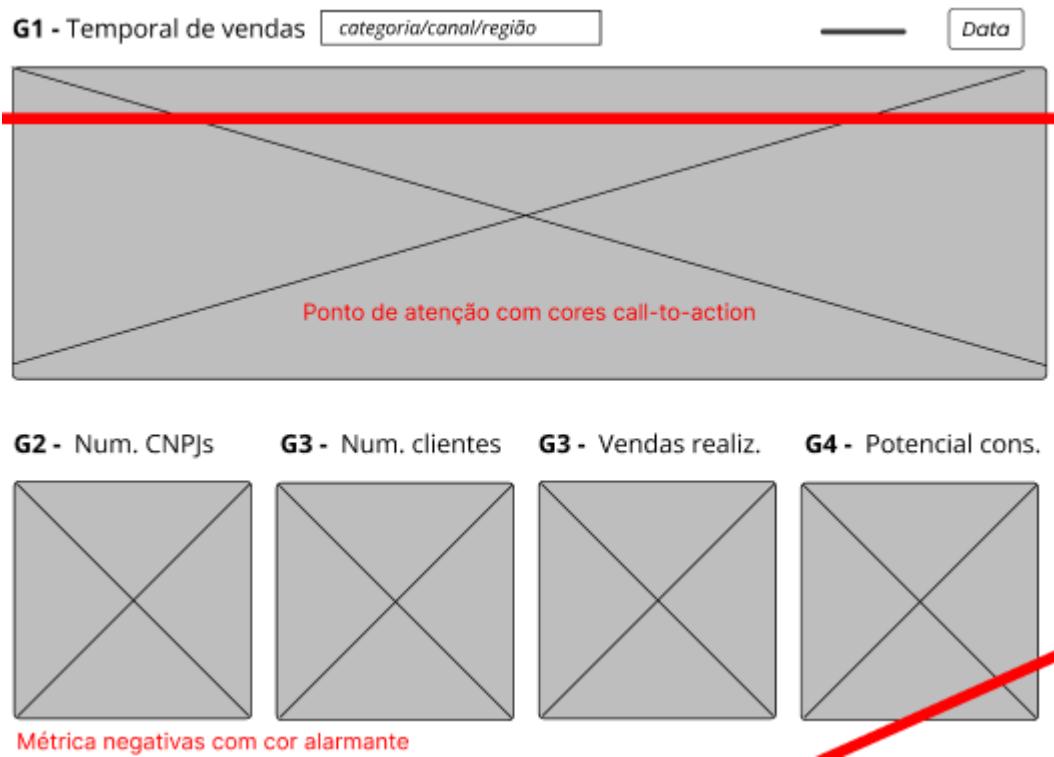
Trouxemos essas escolhas de design através de técnicas que desempenham um papel crucial na criação de experiências digitais fluídas, como a antecipação de ações do usuário, que serão implementadas através de pistas visuais, animações sutis nos botões e gráficos que sugerem o próximo passo lógico:



Figura XX - Hierarquia visual Dashboard

Fonte: Elaborado pelo próprio autor (2023)

A lógica por trás dessas disposições de elementos também está ancorada na hierarquia de informações e na facilidade de acesso às diferentes perspectivas dos dados. Sendo assim destacamos os dados mais importantes e com maior granularidade com mais proximidade do início da hierarquia visual:



Enquanto que, os gráficos que apresentam menor granularidade, estão mais próximos do fim da jornada visual, como por exemplo a visualização de gráficos por setor do mercado. Fizemos essa escolha pois, refletindo na usabilidade do consultor, as métricas micro da empresa - provenientes de categoria, canal e região - podem trazer tomadas de decisão imediatas antes mesmo de se fazer uma análise macro e mais elaboradas pelas métricas de mercado/setor:



Figura XX - Hierarquia visual Dashboard

Fonte: Elaborado pelo próprio autor (2023)

A ferramenta de grid, que proporciona uma estrutura organizada e consistente para o layout de elementos na tela, também foi recrutada. Ao utilizar um grid, é possível alinhar e distribuir elementos de forma equilibrada, garantindo uma aparência profissional e coesa. Além disso, um grid bem elaborado facilita a adaptação do design para diferentes tamanhos de tela e dispositivos, contribuindo para uma experiência responsiva.

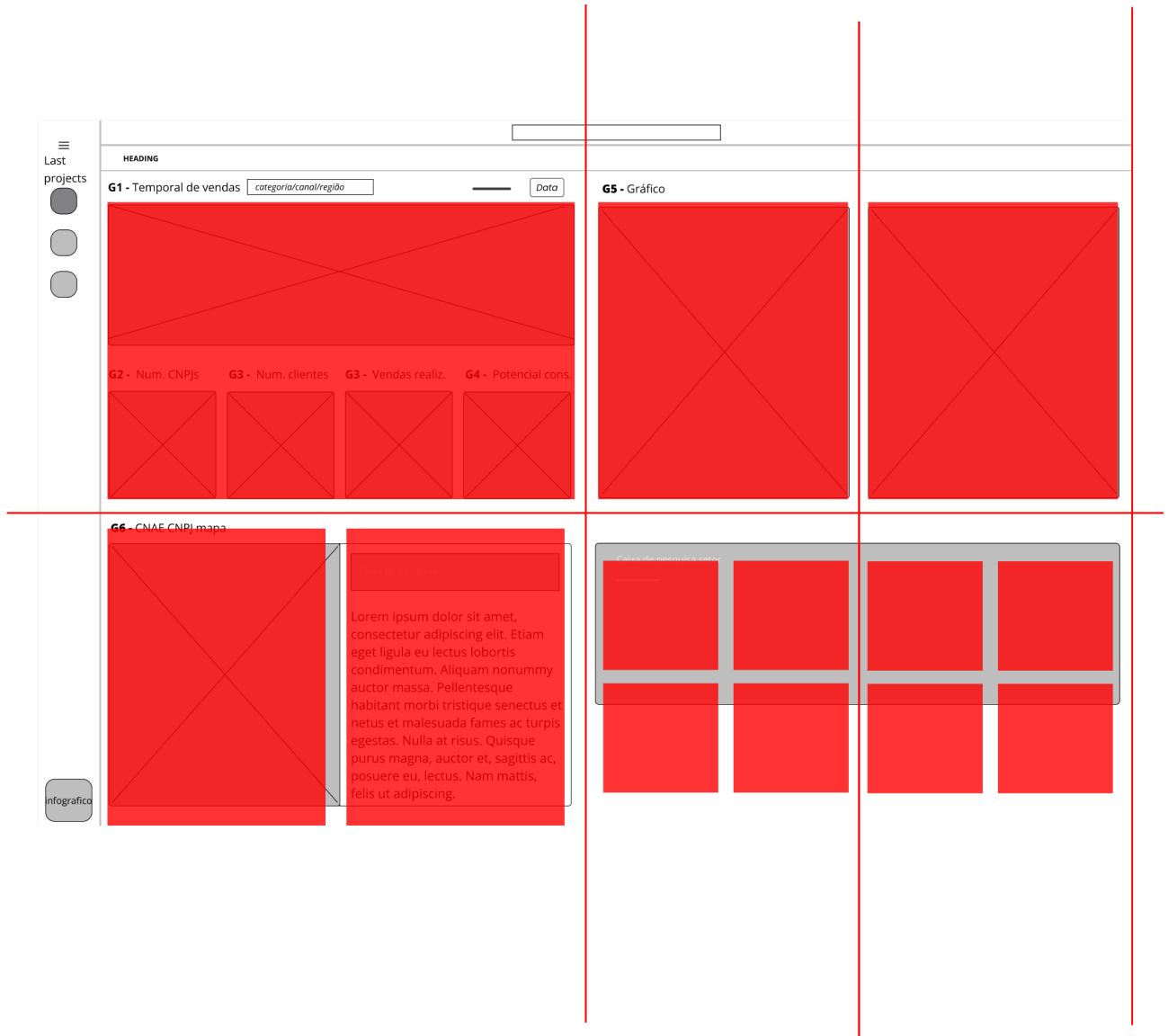


Figura XX - Sistema Grid do Dashboard

Fonte: Elaborado pelo próprio autor (2023)

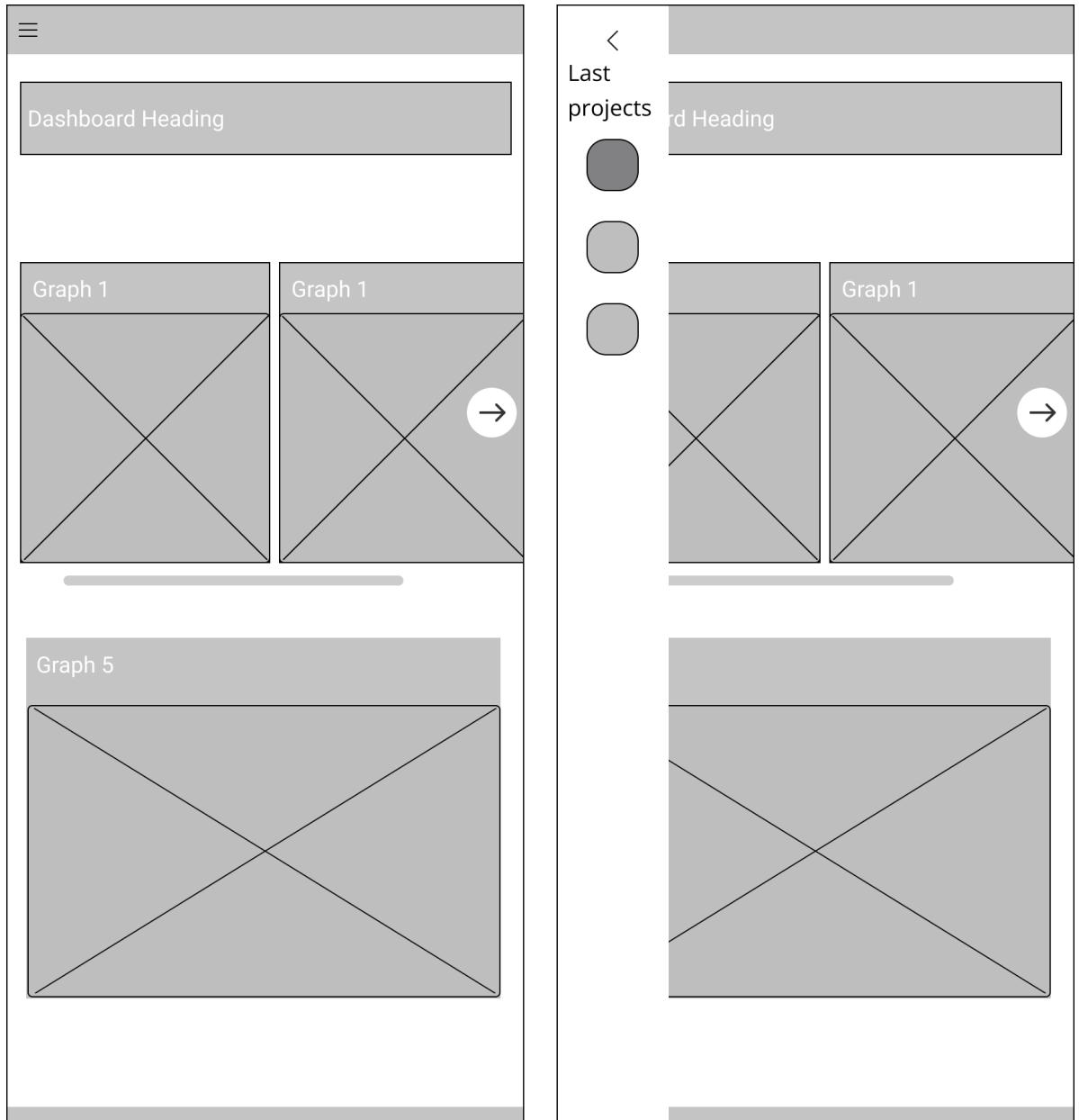


Figura XX - Versão mobile Dashboard

Fonte: Elaborado pelo próprio autor (2023)

5.4 Feedback e Iterações

O processo de feedback e iterações é essencial no desenvolvimento de qualquer projeto de design, pois permite aprimorar a experiência do usuário com base nas percepções e necessidades reais dos usuários. No contexto da tela que estamos discutindo, o feedback coletado desempenhou um papel crucial na evolução do design e nas decisões tomadas.

A partir desse feedback, identificaram-se áreas de melhoria, como a necessidade de uma funcionalidade de pesquisa para facilitar a localização de setores específicos.

Esse elemento foi posteriormente incorporado ao design para proporcionar uma experiência mais personalizada e eficiente.

Além disso, observou-se a importância de tornar as visualizações de dados mais claras e intuitivas. Com base nesse insight, foram feitos ajustes na distribuição do grid e dos mapas, garantindo uma compreensão mais rápida e precisa. O feedback também influenciou a decisão de adicionar a funcionalidade de seleção de opções após a escolha de um setor na caixa de pesquisa. Tal processo proporcionou insights valiosos sobre as preferências e comportamentos dos usuários, permitindo que o design fosse ajustado de maneira mais eficaz às necessidades dos usuários finais.

6. Análise Exploratória e fonte dos dados

6.1 Introdução

Bucket base dos dados

<https://basedosdados.org/dataset/49ace9c8-ae2d-454b-bed9-9b9492a3a642?table=3880670f-eceb-47ec-802b-4579ee62ae3>

Importância dos dados: Os *dados geográficos* brasileiros foram selecionados para entender as demandas regionais.

Formato: CSV

Tamanho: base de dados não operante

Frequência de atualização: trimestralmente (2000-2021)

<https://basedosdados.org/dataset/9fa532fb-5681-4903-b99d-01dc45fd527a?table=4b025d5a-5af0-4fa8-bd04-59de13b378ae>

Importância dos dados: A *Pesquisa Nacional* por Amostra de Domicílios Contínua foi escolhida para entender a situação social das famílias.

Formato: CSV

Tamanho: 14 conjuntos de dados

Frequência de atualização: trimestralmente (2012-2020)

<https://basedosdados.org/dataset/a1b6d2b6-4aa6-47e7-a517-8a21b28b7254?table=7b880731-ffa2-4bde-a290-ae058b3acf51>

Importância dos dados: A *Pesquisa de Orçamentos Familiares* foi escolhida para, complementando os dados geográficos e a PNAD, viabilizar um perfil socioeconômico das famílias brasileiras.

Formato: CSV

Tamanho: 8 conjuntos de dados

Frequência de atualização: trimestralmente (2017-2018)

Bucket dados abertos ibge

<https://dados.gov.br/dados/conjuntos-dados/io-produto-interno-bruto-dos-municípios>

Importância dos dados: O PIB de cada município foi selecionado para adicionar informações sobre concentração de renda.

Formato: HTML; JSON; ODS; XML;

Tamanho: 422 conjuntos de dados

Frequência de atualização: anualmente

<https://dados.gov.br/dados/conjuntos-dados/pc-indice-nacional-de-precos-ao-consumidor-inpc>

Importância dos dados: Índice Nacional de Preços ao Consumidor para demonstrar a variância dos preços.

Formato: HTML; JSON; ODS; XML;

Tamanho: 422 conjuntos de dados

Frequência de atualização: anualmente

Bucket POF IBGE

<https://www.ibge.gov.br/estatisticas/sociais/saude/24786-pesquisa-de-orcamentos-familiares-2.html?=&t=downloads>

Importância dos dados: Pesquisa de Orçamentos Familiares dividido por décadas.

Formato: TXT

Tamanho: 8 conjuntos de dados

Frequência de atualização: anualmente (1987-1988; 1995-1996; 2002-2003; 2008-2009; 2017-2018)

Bucket Microdados RAIS e CAGED

<https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/estatisticas-trabalho/microdados-rais-e-caged>

Importância dos dados: As empresas fornecem a Relação Anual de Informações Sociais e, junto com o Cadastro de Geral de Empregados e Desempregados, são fornecidas informações sobre a população economicamente ativa.

Formato: Indisponível

Tamanho: Indisponível

Frequência de atualização: Indisponível

Bucket Receita Federal

<https://dados.gov.br/dados/conjuntos-dados/resultado-da-arrecadacao>

Importância dos dados: Resultado da arrecadação com objetivo de sintetizar informações da economia brasileira.

Formato: CSV; PDF

Tamanho: 134 conjuntos de dados

Frequência de atualização: diariamente

<https://dados.gov.br/dados/conjuntos-dados/repasses-da-arrecadacao-federal>

Importância dos dados: Repasses da união aos estados.

Formato: CSV; PDF; XLSX

Tamanho: 134 conjuntos de dados

Frequência de atualização: diariamente

<https://dados.gov.br/dados/organizacoes/visualizar/ministerio-da-fazenda>

Importância dos dados: Dados relativos a título e tesouro.

Formato: CSV; PDF; XLSX; JSON

Tamanho: 134 conjuntos de dados

Frequência de atualização: diariamente

Bucket INEP

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/ana>

Importância dos dados: A Avaliação Nacional de Alfabetização adiciona informações sobre a educação básica da população.

Formato: CSV

Tamanho: 3,300 KB

Frequência de atualização: anualmente

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>

Importância dos dados: Evolução anual da educação brasileira (1995-2022)

Formato: CSV

Tamanho: aproximadamente 8 GB

Frequência de atualização: anualmente (1995-2022)

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/encceja>

Importância dos dados: Microdados do Exame Nacional para Certificação de Competências de Jovens e Adultos

Formato: CSV

Tamanho: 0.6 GB

Frequência de atualização: anualmente (2014, 2017-2020, 2022)

Arquivo:

<https://drive.google.com/drive/folders/1GDP0qcn8XWYdv0c1w46EK2eV013yXShi?usp=sharing>

Bucket SUS

https://opendatasus.saude.gov.br/dataset/cnes-cadastro-nacional-de-estabelecimentos-de-saud_e

Importância dos dados: Cadastro Nacional de Estabelecimentos de Saúde

Formato: JSON

Tamanho: 832 MB

Frequência de atualização: diariamente

Bucket POF_parceiro

https://drive.google.com/drive/folders/1kB92-Q_pDSIZ4YTxGy3ygxpylKA19mwJ

Importância dos dados: Pesquisa de Orçamentos Familiares disponibilizado pelo parceiro

Formato: CSV

Tamanho: 900 MB

Frequência de atualização: anualmente

Bucket cnpj

<https://drive.google.com/drive/folders/1hxU0AdkgvT23FRGNp4htNs45WC5mYC1S>

Importância dos dados: CNPJ's das empresas, tanto as que compram quanto as que vendem alimentos

Formato: CSV

Tamanho: 4,582 GB

Frequência de atualização: anualmente

6.2 Método

7. Arquitetura Macro

Para a execução do projeto foi necessário a definição de uma arquitetura cloud que será guia para a construção do ambiente na cloud da AWS e definição das tecnologias a serem usadas. O esquema abaixo corresponde a arquitetura realizada na Sprint 1 para apresentação e alinhamento de expectativas com os parceiros de projeto, assim também é esperado uma dinâmica de iteração durante o decorrer das Sprints.

Imagen 1: [Arquitetura](#)

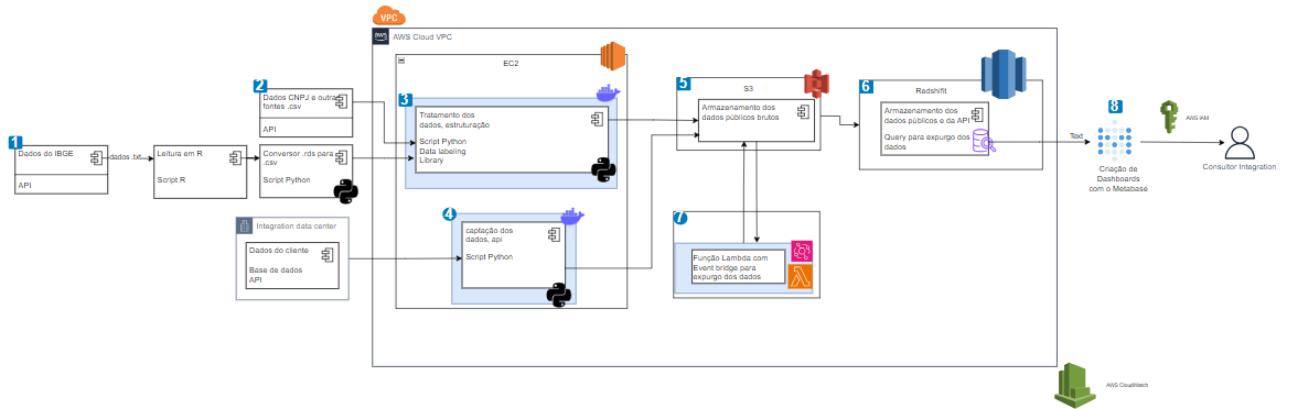


Figura XX - Fonte: Elaborado pelo próprio autor (2023)

Abaixo se encontra uma análise da arquitetura construída:

7.1 Identificação dos dados:

7.1.1 Dados públicos:

São fornecidos pelos canais de pesquisa do governo, tendo como mais frequente o IBGE, através de arquivos .txt em uma pasta no início do projeto. De partida, os dados não apresentavam-se estruturados, sendo fornecido um script em R para leitura desses dados e a criação de tabelas que possibilitam a compreensão. Após a leitura os dados serão convertidos para .rds e posteriormente para .csv e passados para o EC2. Nos dados públicos também se encaixam os dados referentes aos CNPJ no Brasil, que foram fornecidos diretamente em .csv.

Conteúdo: público, dados demográficos coletados pelo governo, que englobam fatores como condição financeira dos moradores de cada região, hábitos de consumos alimentares, despesas, aluguel dentre outros, organizados por fatores como faixa etária, faixa de renda e região. Os dados de Pesquisa de Orçamentos Familiares (POF) contemplam os seguintes tópicos:

- Primeiros Resultados, que apresenta informações sobre despesas e rendimentos das famílias.
- Avaliação nutricional da disponibilidade domiciliar de alimentos no Brasil.
- Análise do consumo alimentar pessoal no Brasil.
- Análise da Segurança Alimentar no Brasil.
- Perfil das Despesas no Brasil: indicadores selecionados.
- Perfil das Despesas no Brasil: indicadores selecionados de alimentação, transporte, lazer e inclusão financeira.
- Perfil das Despesas no Brasil: indicadores de qualidade de vida.
- Evolução dos indicadores de qualidade de vida no Brasil com base na Pesquisa de Orçamentos Familiares.

Tabelas: Os dados se distribuem em tabelas de acordo com o seu conteúdo, sendo as fornecidas pelo parceiro:

Tabela	Descrição
Aluguel Estimado	Pesquisa sobre os custos estimados de aluguel no Brasil.
Caderneta - Coletiva	Levantamento de informações financeiras coletivas em cadernetas.
Características - Dieta	Estudo das características da dieta alimentar da população.
Condições - Vida	Análise das condições de vida no país.
Consumo - Alimentar	Investigação sobre o consumo de alimentos na sociedade.
Despesa - Coletiva	Coleta de dados sobre despesas coletivas.
Despesa - Individual	Coleta de dados sobre despesas individuais.
Domicílio	Levantamento de informações relacionadas a residências.
Inventário	Pesquisa sobre o inventário de bens e recursos.
Morador - Qualidade - Vida	Avaliação da qualidade de vida dos moradores.
Morador	Coleta de dados sobre os habitantes de uma área específica.
Outros - Rendimentos	Pesquisa de rendimentos não especificados.
Rendimento - Trabalho	Estudo dos rendimentos provenientes do trabalho.
Restrição - Produtos - Serviços - Saúde	Avaliação das restrições no acesso a produtos e serviços de saúde.
Serviço não monetário - POF2	Pesquisa sobre serviços não monetários na Pesquisa de Orçamento Familiar (POF) 2.
Serviço não monetário - POF4	Pesquisa sobre serviços não monetários na Pesquisa de Orçamento Familiar (POF) 4.

7.1.2 Dados do cliente:

Os dados referentes ao cliente serão fornecidos através de uma API hospedada pela Integration, esses dados contém informações privadas sobre os clientes da Integration, portanto, são de natureza sigilosa, não permitindo ficar em posse da Integration após o projeto. Esses dados são obtidos através de queries realizadas pela API da Integration.

Conteúdo: Sigiloso, o conteúdo desses dados engloba informações específicas sobre o parceiro, compreendendo dados como receita gerada por um produto, receita regional, número de vendas, lista de produtos, CNPJ de clientes dentre outros. Esses dados devem variar de cliente para cliente e refletem sua situação no mercado. A solução precisa estar preparada para receber diferentes tipos de dados e estruturas, pois eles podem variar de cliente para cliente.

7.2 Gestão de dados:

7.2.1 Dados públicos:

Ingestão: Os dados são coletados por canais públicos como do Instituto Brasileiro de Geografia e Estatística - IBGE e outros órgãos públicos. Através de uma API, fornecida pelo governo, podem-se ser obtidos em formato .txt, sendo necessário passar por scripts em R para transformá-los em tabelas legíveis e distinguíveis. Esses dados passarão por uma pipeline em python na EC2, juntamente com os dados referentes aos CNPJ, e depois carregados em sua totalidade no S3. Após isso, os dados relevantes para os projetos atuais da empresa devem ser transferidos para o Redshift, mantendo os dados mais importantes lá. Os quais posteriormente serão consumidos no EC2 por um código python, o qual será processado com o sparky, e também será parte do cubo de dados.

frequência: A atualização desses dados deve ser feita anualmente, acompanhando a velocidade a qual os censos são atualizados.

quantidade: Espera-se que quando atualizados será em grande quantidade, pois a frequência de atualização não será tão alta, e os dados do governo englobam milhões de pessoas.

7.2.2 Dados privados:

Já para os dados privados os dados serão ingeridos a partir da Api que fornece os dados do cliente em questão para cada projeto. A Api será consultada por um código

python rodando no EC2, e posteriormente carregado, em um código python, também na EC2, utilizando sparky para uma manipulação dos dados. Eles também serão utilizados para a criação de um cubo de dados.

frequência: A atualização desses dados deve ser feita diariamente, gerenciada por um airflow para essa execução, a fim de manter os dados sempre atualizados em relação aos dados fornecidos na api do parceiro.

quantidade: A quantidade de dados, embora grande, será recebida em um quantidade média, pois serão atualizados com uma frequência maior.

7.3 Seleção dos serviços AWS:

Tratamento/processamento de dados:

EC2: realizaremos o tratamento e processamento dos dados pela máquina virtual do Amazon EC2 (Elastic Compute Cloud), através de códigos python. O EC2 é um serviço de nuvem da Amazon que oferece servidores virtuais para hospedar os recursos computacionais para a pipeline de tratamento.

Serão no total quatro códigos alocados na EC2:

Tratamento de dados públicos: Esse código receberá os dados públicos, os processará e os enviará para o AWS S3.

Consulta da api/Redshift: Esse código será responsável por consultar a api do parceiro e os códigos guardados no Redshift, criar um cubo de dados com essas informações e, passar as informações para o código de tratamento/transformação de dados.

Código de tratamento/ transformação de dados com sparky: Esse código será responsável pelo processamento dos tanto do cliente tanto dos dados públicos, visando alcançar as informações solicitadas pelo parceiro.

Código para o infográfico: Esse código será responsável por criar infográficos com base nos dados fornecidos, com objetivos de fornecer possíveis insights sobre os dados.

Armazenamento:

S3: Por aspectos do pipeline de big data como volume e custos, será utilizado o , serviço que permite o armazenamento de dados de forma flexível e não estruturada. Nele se espera comportar todos os dados públicos com alto volume histórico

de datação prezando pela economia em espaço e processamento alcançáveis por uma base de dados não relacional.

Redshift: Posteriormente, após a seleção dos dados de interesse para as análises de mercado e negócios, será utilizado o AWS Redshift para hospedar um data warehouse, o qual receberá os dados mais relevantes que estavam no S3. Este passo restringirá a presença no Redshift apenas aos dados essenciais para a elaboração de tabelas e informações de negócios necessárias para a próxima etapa de processamento e análise.

Segurança de dados:

VPC: Possibilita a criação de secções na nuvem da aws, sendo possível controlar o acesso aos serviços da AWS utilizados, criar restrições de acesso e regras de segurança. Sendo possível garantir uma segurança maior à solução.

Monitoramento:

Amazon Cloudwatch: Será utilizado o AWS Cloud Watch para o monitoramento dos serviços mencionados anteriormente, EC2 e Redshift. Visualizando informações sobre eles e o status em tempo real.

7.4 Fluxo dos dados:

Os dados possuem diferentes fluxos, mas se conectam futuramente no EC2, segue abaixo os fluxos dos dados até se juntarem:

Dados públicos: Os dados públicos serão retirados dos sites do governo em formato .txt, após isso os dados terão que passar por um Script R para gerar arquivos .rds a partir deles e após isso por um script python para conversão em .csv. Após isso eles serão inseridos no EC2, em um notebook que irá tratar os dados e os enviar para o S3 em sua totalidade. Os dados mais relevantes atualmente para empresa serão selecionados e enviados para o Redshift e por fim enviados de volta para o EC2 para criação do cubo de dados.

Dados privados: Os dados privados serão solicitados diretamente na api fornecida pelo parceiro, esse consumo será realizado em código python localizado na EC2 e utilizados para criação do cubo de dados.

Cubo de Dados: Após a formação inicial do cubo de dados os dados passarão por outro código python no EC2, sendo processados e agregados de forma a gerar as informações que o cliente deseja para gerar um relatório. E por fim serão enviados para outro código python no EC2, o qual irá gerar infográficos a partir desses dados.

7.5 Infográfico:

O infográfico, construído com base nos dados previamente processados e organizados no cubo de dados, será desenvolvido utilizando a Metabase. A escolha desta ferramenta deve-se à sua interface de usuário intuitiva e à sua disponibilidade como uma solução sem custos, o que a torna particularmente adequada e conveniente para as necessidades e restrições do projeto.

7.6 Segurança:

Como mencionado anteriormente, uma das ferramentas utilizadas para garantir a segurança do projeto foi o Amazon VPC, que será responsável pelo controle de acesso, criação de restrições e regras de segurança no ambiente da AWS.

Mas além dos cuidados com a segurança da AWS, também foram tomados cuidado com os dados, uma vez que apenas dados públicos serão salvos na base de dados, evitando assim que os dados privados do parceiro obtidos através da api tenham risco de vazamento.

7.7 Monitoramento e gerenciamento:

Como mencionado anteriormente na parte de serviços será utilizado o Amazon CloudWatch para o monitoramento da solução. Dentre os tópicos monitorados vale destacar:

Gasto de recursos: Para garantir que as ferramentas da AWS estejam utilizando os recursos esperados, de forma a evitar gastos indesejados.

Disponibilidade: Verificar a disponibilidade dos serviços, de forma a garantir o funcionamento da solução.

Informações: Monitoramento do que está ocorrendo nas ferramentas da solução, de forma a garantir que o fluxo esteja ocorrendo como planejado.

8. Data lake

8.1 Introdução

Um Data Lake é um sistema centralizado que permite armazenar grandes volumes de dados em seu formato natural, seja estruturado, semi-estruturado ou não estruturado. A ideia é que você possa despejar dados de diferentes fontes dentro deste lago e eles estarão disponíveis para análise e processamento com grande flexibilidade.

No nosso caso, o objetivo de criarmos uma pipeline de dados para a Integration, faz com que exista a necessidade da criação de um data lake, uma vez que estamos trabalhando com um grande volume de dados e de diferentes fontes. Para isso utilizaremos o S3.

8.2 S3

O S3, significa Simple Storage Solution, e a escolhemos por várias razões. Lançado em 2006, ele oferece um armazenamento de objetos dentro de uma estrutura de pastas que é extremamente eficaz em custo, começando em apenas \$0,023 por gigabyte. Além disso, quanto mais você armazena, mais o custo diminui, sem limitações de capacidade.

Ele proporciona uma maneira barata e confiável de armazenar objetos com acesso de baixa latência e alta capacidade de transferência através do conteúdo do seu bucket. Um Bucket é um container da AWS S3 para armazenamento de diversas opções de arquivos. Outro ponto forte é a sua integração com serviços como SNS, SQS e Lambda, possibilitando aplicações poderosas orientadas por eventos, podendo utilizar Lambda através de um trigger. Além disso, o S3 oferece mecanismos para transferir dados antigos para armazenamento de longo prazo, reduzindo ainda mais os custos.

8.3 Buckets

O processo para utilizar o S3 é simples. Primeiro, cria-se um bucket. Depois, os arquivos são carregados para este bucket, com configurações personalizáveis (as quais explicamos a configuração abaixo). É importante notar que, embora o S3 utilize o termo "pastas", ele opera com uma estrutura de "flat file structure". Isso significa que não é como as pastas do Windows; o nome do arquivo é anexado ao nome do arquivo como um prefixo, o que otimiza a organização e o acesso aos dados.

Os buckets como dito anteriormente é um instrumento da AWS S3 que se assemelha com um container do docker. Nele você pode fazer o upload de arquivos, como JSON, PNG, HTML, CSV, etc, no caso de nosso projeto o formato escolhido foi o .csv. Se acessado na nuvem através de seu link de acesso, que pode ser definido geograficamente na própria configuração do Bucket.

Para isso criamos 10 buckets que seguem a seguinte configuração.

8.3.1 Configuração dos Buckets

Passo 1: Configuração inicial

Nome do bucket

nome-exemplo

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

Região da AWS

América do Sul (São Paulo) sa-east-1



Para começar a configuração damos o nome do bucket, (ele tem que ser único nos servidores da AWS) e a localização. Por estarmos fazendo na AWS Workbench não pudemos hospedar nosso bucket em São Paulo, mas esse seria o recomendado.

Passo 2: ACL

Propriedade de objeto [Informações](#)

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

ACLs desabilitadas (recomendado)

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

ACLs habilitadas

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

Desabilitamos ACLs no bucket S3 para simplificar a gestão de permissões.

Passo 3: Bloqueios e privacidade

Configurações de bloqueio do acesso público deste bucket

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

Bloquear todo o acesso público

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

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

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

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

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

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

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

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

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

Para a prova de conceito, optamos por bloquear todo o acesso público ao bucket S3, garantindo segurança e simplicidade na gestão de permissões. A Integration, em um contexto de produção, deve considerar políticas específicas, possivelmente reabilitando ACLs para acesso detalhado, conforme a necessidade do cliente ou a sensibilidade dos dados.

Passo 4: Versionamento de bucket

Versionamento de bucket

O versionamento é um meio de manter múltiplas variantes de um objeto no mesmo bucket. Você pode usar o versionamento para preservar, recuperar e restaurar todas as versões de cada objeto armazenado no bucket do Amazon S3. Com o versionamento, você pode recuperar facilmente ações não intencionais do usuário e falhas da aplicação. [Saiba mais](#)

Versionamento de bucket

- Desativar
- Ativar

Decidimos não ativar o versionamento do *bucket*, considerando a natureza dos dados que estamos manuseando. Como são dados atualizados anualmente, não há necessidade de manter múltiplas versões ao longo do ano, o que poderia resultar em **custos desnecessários**.

Passo 5: Criptografia

Criptografia padrão [Informações](#)

A criptografia no lado do servidor é aplicada automaticamente a novos objetos armazenados nesse bucket.

Tipo de criptografia | [Informações](#)

- Criptografia do lado do servidor com chaves gerenciadas do Amazon S3 (SSE-S3)
- Criptografia do lado do servidor com chaves do AWS Key Management Service (SSE-KMS)
- Criptografia de duas camadas no lado do servidor com chaves do AWS Key Management Service (DSSE-KMS)
Proteja seus objetos com duas camadas separadas de criptografia. Para obter detalhes sobre a especificação, consulte os preços do DSSE-KMS na guia [Armazenamento da página de preços do Amazon S3](#).

Chave do bucket

O uso de uma chave de bucket do S3 para SSE-KMS reduz os custos de criptografia ao diminuir as chamadas para o AWS KMS. As chaves de bucket do S3 não são compatíveis com o DSSE-KMS. [Saiba mais](#)

- Desativar
- Ativar

Para essa prova de conceito, optamos pela criptografia do lado do servidor fornecida pela própria AWS S3, a SSE-S3. Isso nos dá segurança de forma prática, sem a necessidade de gerenciar as chaves de criptografia manualmente, o que seria necessário se escolhêssemos outras opções como SSE-KMS ou DSSE-KMS.

Para a equipe da Integration, vale apontar que a escolha pelo SSE-S3 é totalmente adequada para essa fase inicial, onde estamos tratando de uma prova de conceito e de dados públicos. No entanto, quando o projeto tomar forma e começar a lidar com um volume maior de dados ou informações mais sensíveis, aí sim faria sentido pensar em algo mais robusto como o SSE-KMS ou DSSE-KMS. Essas opções trazem benefícios adicionais, como um maior controle e auditoria sobre quem está acessando as chaves de criptografia.

9. Cubo de dados e data warehouse

9.1 Introdução:

A criação de cubo de dados e de um data warehouse desempenha papéis cruciais em projetos de Big Data com o objetivo de estabelecer um pipeline de dados eficiente. O cubo de dados oferece a capacidade de agrregar dados multidimensionalmente, melhorando o desempenho de consultas e permitindo análises detalhadas. Ele facilita a exploração de informações em diversas perspectivas e integra-se facilmente a ferramentas de Business Intelligence. Por sua vez, o data warehouse atua como um repositório centralizado, consolidando dados de várias fontes e garantindo qualidade e consistência por meio da limpeza e transformação. No contexto de um pipeline de dados, como o do nosso projeto, essa combinação proporciona agilidade na manipulação, movimentação e preparação dos dados, essenciais para análises rápidas e eficientes em ambientes de Big Data.

A criação de um cubo de dados em um data warehouse envolve entender os requisitos de negócios, extrair, transformar e carregar dados (ETL), modelar dimensões, construir o cubo e implementá-lo no data warehouse. Os usuários finais acessam o cubo através de ferramentas de Business Intelligence para análises. A manutenção contínua é essencial para ajustar o modelo conforme necessário, garantindo que os dados estejam sempre atualizados e atendam aos requisitos em evolução. Este processo fornece uma base robusta para análises eficazes.

9.2 Fonte de dados:

Os dados escolhidos para construção do cubo de dados são formados por dois tipos.

Dados públicos: constituídos dos dados trabalhados anteriormente na exploratória, e tiveram suas especificações, como volume, tipo e formato trabalhados em nossa análise exploratória. Esses dados serão retirados de fontes governamentais e após passarem pelo “script para subir dados no s3” serão enviados em formato csv para buckets no AWS S3 de acordo com suas respectivas fontes.

Dados privados(API): Esses dados são consumidos diretamente da API do parceiro, constituída de 3 tabelas Category, Establishment e Sale. E esperados que esses dados sejam consumidos periodicamente por um script python e salvos no formato csv em nosso S3 em bucket separado exclusivamente para os dados dessa API.

Mas mesmo sendo dados provenientes de fontes diferentes todos os dados consumidos pelo Redshift foram concentrados na mesma ferramenta o AWS S3, sendo assim possível extrair todos os dados de uma mesma fonte.

9.3 ETL:

O processo de ETL é de extrema importância para alimentação de nossa data warehouse, uma vez que ele garante que os dados serão padronizados e armazenados de forma correta. A ETL consiste de três etapas, que serão indicadas a seguir:

Extração: Para essa etapa nossa solução realizou a extração de todos os dados trabalhados no tópico “Fonte de dados” de uma única fonte, os buckets localizados no AWS S3. para essa extração foi realizada diretamente da ferramenta do Redshift, sendo destinado a diferentes tabelas para cada tópico encontrado em nossos buckets.

Tratamento: Após a extração dos dados foi realizado o tratamento e a padronização dos mesmos. O tratamento foi feito visando formatar as tabelas em um formato padrão que possua as seguintes colunas:

data_ingestão: Visando guardar o horário de ingestão dos dados no formato TIMESTAMP, de forma a possibilitar um controle temporal dos dados armazenados.

tag: Visando guardar o tipo/fonte desses dados de forma a ser possível os identificar de acordo com seu tipo, no formato de string.

value: Coluna responsável por guardar os JSON de cada linha dos dados consumidos no S3.

use: Coluna que indica qual é a utilização do dados no formato de string.

carregamento dos dados: Para essa etapa foi utilizada a ferramenta AWS Redshift para o carregamento e armazenamento dos dados, após a sua coleta no S3, os dados serão destinados a diferentes tabelas para cada tópico encontrado em nossos buckets, sendo armazenados no padrão mencionado no ETL.

9.4 serviço escolhido, Redshift:

Para a escolha desse serviço, foram analisadas tanto as nossas necessidades como as limitações e capacidades dos serviços analisados.

Primeiramente vale a pena ressaltar a necessidade dos dados serem disponibilizados através de um OLAP devido a sua capacidade de processar e armazenar grandes volumes de dados. Sendo a solução escolhida para lidar com as diferentes fontes de dados e os grande volume de dados que teriam que ser armazenados e analisados frequentemente.

Sendo assim, foi realizado o levantamento dos serviços disponíveis no ambiente da AWS, sendo cotados serviços como RDS e DynamoDB. Mas esses serviços foram eliminados de nossa seleção, uma vez que não possuíam a otimização adequada para armazenamento e disponibilização de um OLAP.

Após analisar os serviços disponíveis no ambiente AWS, o serviço escolhido foi o AWS Redshift, uma vez que este é a forma de armazenamento com a maior compatibilidade com cubos de dados OLAP. Sendo possível atender diversas necessidades que nós tínhamos com o projeto, como:

Armazenamento Colunar: O Redshift armazena dados em formato colunar, possibilitando consultas analíticas otimizadas, e a agregação e a análise de grandes volumes de dados de forma rápida e eficiente.

Computação distribuída: O Redshift permite que você distribua e paralelize consultas em várias máquinas, acelerando assim as consultas realizadas.

Consultas SQL: O Redshift permite a criação e execução de consultas complexas.

Integração com Ferramentas de BI: O Redshift pode ser consumido várias ferramentas de Business Intelligence para criação de nosso infográfico.

Escalabilidade: O Redshift é um serviço que fornece alta escalabilidade, possibilitando o crescimento da solução.

Portanto devido tanto a seu suporte ao olap, quanto aos benefícios mencionados o serviço escolhido foi o AWS Redshift

9.5 Estrutura e configuração do Redshift:

Configuração do Redshift:

Para criação de nosso datawarehouse, como mencionado no tópico de escolha do serviço, foi utilizada a ferramenta AWS Redshift, sendo realizadas diversas configurações para seu funcionamento e otimização, com intuito de maximizar a eficiência pelo menor custo o possível. Segue abaixo as etapas realizadas para criação de nosso Redshift:

1- definição do nome, e forma de configuração:

The screenshot shows the AWS Redshift Serverless setup interface. At the top, there's a navigation bar with the AWS logo, a search bar containing '(Opção+S)', and account information for 'Norte da Virgínia' and 'lucasbritto'. Below the navigation, the main title is 'Primeiros passos com o Amazon Redshift Serverless' with a 'Informações' link. A sub-section titled 'Configuração' contains two options: 'Usar configurações padrão' (selected) and 'Personalizar configurações'. The 'Personalizar configurações' section is described as allowing users to 'Personalize suas configurações de acordo com suas necessidades específicas'. Below this, there's a 'Espaço para nome' section with a 'Informações' link, describing namespaces as collections of objects and users. It shows a 'Namespace de destino' input field containing 'redshift-bigD'. Under 'Nome e senha do banco de dados', there's a 'Nome do banco de dados' input field containing 'dev'. At the bottom of the page, there are links for 'CloudShell', 'Comentários', and legal notices: '© 2023, Amazon Web Services, Inc. ou suas afiliadas.', 'Privacidade', 'Termos', and 'Preferências de cookies'.

2- Definição de função do IAM com permissões para acesso e modificação do Redshift:

Funções do IAM associadas (0)

- Definir padrão
- Gerenciar funções do IAM
- Associar funções do IAM
- Criar função do IAM**
- Remover funções do IAM

Não há recursos

Nenhuma função do IAM associada

Assoclar função do IAM

Segurança e criptografia

Seus dados são criptografados por padrão com uma chave de propriedade da AWS. Para escolher uma chave diferente, personalize suas configurações de criptografia.

3- Definição de nome do grupo de trabalho, para controle do ambiente:

4- definição de RPU, para definição da capacidade de computação do Redshift:

Grupo de trabalho Informações

Grupo de trabalho é uma coleção de recursos de computação a partir dos quais um endpoint é criado. As propriedades de computação incluem configurações de rede e segurança.

Nome do grupo de trabalho

Este é um nome exclusivo que define o grupo de trabalho.

O nome deve ter de 3 a 64 caracteres. Os caracteres válidos são az (somente letras minúsculas), 0-9 (números) e - (hifen).

Capacidade básica em unidades de processamento (RPUs) do Redshift

Defina a capacidade básica usada para processar sua workload. Para melhorar a performance da consulta, aumente seu valor de RPU.

Capacidade básica de RPU

A capacidade básica de RPU é definida como automática por padrão, que é igual a 128 RPUs. Para alterar a capacidade básica de RPU, escolha outro valor da lista.

O intervalo deve ser de 8 a 512 em incrementos de 8.

Rede e segurança

Nuvem privada virtual (VPC)

Essa VPC define o ambiente de redes virtual para esse banco de dados.

Grupos de segurança da VPC

Esse grupo de segurança da VPC define quais sub-redes e intervalos de IP podem ser usados na VPC.

5- Definições de VPC, para o controle de acesso e segurança da data warehouse:

Amazon Redshift | us-east-1

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#serverless-setup

Serviços | Procurar | [Opção+S]

128

O intervalo deve ser de 8 a 512 em incrementos de 8.

Rede e segurança

Nuvem privada virtual (VPC)

Esse VPC define o ambiente de redes virtual para esse banco de dados.

vpc-0ec9e5efdc417a856

Grupos de segurança da VPC

Esse grupo de segurança da VPC define quais sub-redes e intervalos de IP podem ser usados na VPC.

Escolha um ou mais grupos de segurança

sg-030f8e1fe7558f240

Sub-rede

A sub-rede na VPC escolhida associada ao banco de dados especificado.

Escolha três ou mais IDs de sub-rede

sub-rede-04cce91d7d11e2e8c × sub-rede-0e3087c2b41577541 ×
sub-rede-059cd8d103cde975b × sub-rede-00bb9c8aa8b692210 ×
sub-rede-06c6528092876aafa × sub-rede-0c0fb3546f9070bee ×

Roteamento aprimorado da VPC

Ativar essa opção encaminha o tráfego de rede entre o banco de dados sem servidor e os repositórios de dados por meio de uma VPC em vez da Internet.

Ativar roteamento aprimorado da VPC

CANCELAR SALVAR CONFIGURAÇÃO

CloudShell Comentários © 2023, Amazon Web Services, Inc. ou suas afiliadas. Privacidade Termos Preferências de cookies

6- Salvamento da configuração, e aguardar tempo de criação:

Amazon Redshift | us-east-1

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#serverless-setup

128

O intervalo deve ser de 8 a 512 em incrementos de 8.

Rede e segurança

Nuvem privada virtual (VPC)

Esse VPC define o ambiente de redes virtual para esse banco de dados.

vpc-0ec9e5efdc417a856

Grupos de segurança da VPC

Esse grupo de segurança da VPC define quais sub-redes e intervalos de IP podem ser usados na VPC.

Escolha um ou mais grupos de segurança

sg-030f8e1fe7558f240

Sub-rede

A sub-rede na VPC escolhida associada ao banco de dados especificado.

Escolha três ou mais IDs de sub-rede

sub-rede-04cce91d7d11e2e8c × sub-rede-0e3087c2b41577541 ×
sub-rede-059cd8d103cde975b × sub-rede-00bb9c8aa8b692210 ×
sub-rede-06c6528092876aafa × sub-rede-0c0fb3546f9070bee ×

Roteamento aprimorado da VPC

Ativar essa opção encaminha o tráfego de rede entre o banco de dados sem servidor e os repositórios de dados por meio de uma VPC em vez da Internet.

Ativar roteamento aprimorado da VPC

A conclusão pode levar alguns minutos. Depois de concluir a configuração, você poderá trabalhar com seus dados.

Configurar o Amazon Redshift Serverless

Gerenciar dados e computação orientada por machine learning

Fornça constantemente operações simplificadas e de alta performance para as workloads até mesmo mais exigentes e voláteis com escalabilidade inteligente e automática em segundos.

Continuar

CANCELAR SALVAR CONFIGURAÇÃO

CloudShell Comentários © 2023, Amazon Web Services, Inc. ou suas afiliadas. Privacidade Termos Preferências de cookies

Criação de Tabelas:

A criação das tabelas foi realizada com base nos buckets presentes no S3, que constituem nossa fonte de dados, retratada no tópico anterior. Para criação das tabelas foram considerados os tópicos presentes em cada bucket do S3, sendo criada uma tabela diferente para cada tópico e juntando os com colunas iguais em uma única tabela.

Para criação dessa tabelas foi utilizado o próprio Redshift, no qual foram fornecidas primeiramente informações de localização dos dados no S3, o servidor em que estão hospedados e o formato em que os dados estão. Segue abaixo um exemplo dessa configuração:

Load data

Data source

Load from S3 bucket Load from local file

S3 URI:

Region: sa-east-1 ▾ Manifest file

File format: CSV ▾ File options ➔ No compression ▾

Delimiter character: ,
Specifies the single ASCII character that is used to separate fields in the input file, such as a pipe character ('|'), a comma (,), or a tab (\t).

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

Advanced settings Data conversion parameters ➔ Load operations ➔

Após isso também é necessário uma verificação dos tipos de variáveis e limitações da tabela, visando garantir que todos os dados estejam no formato e padrão desejado.

Juntamente com a definição dos nomes, localização de armazenamento e a permissão de acesso da tabela, que também devem ser definidos nessa etapa:

The screenshot shows the 'Load data' interface in the AWS Redshift console. The 'Load new table' option is selected. The 'Cluster or workgroup' is set to 'cubointegration', 'Database' to 'dev', 'Schema' to 'public', and 'Table' to 'sebrae_empregados'. The 'IAM role' dropdown contains 'arn:aws:iam::392330996159:role/iam_redshift'. The 'Columns' tab is active, showing a table with four columns: 'UF' (VARCHAR), 'Mapa de emp...' (INTEGER), 'população' (INTEGER), and 'Relação emp...' (VARCHAR). The 'Column options' panel on the right shows 'Default value' set to 'No default value', 'Size' with a dropdown, and a 'Primary key' checkbox. Buttons at the bottom include 'Back', 'Restore to defaults', 'Cancel', and 'Create table'.

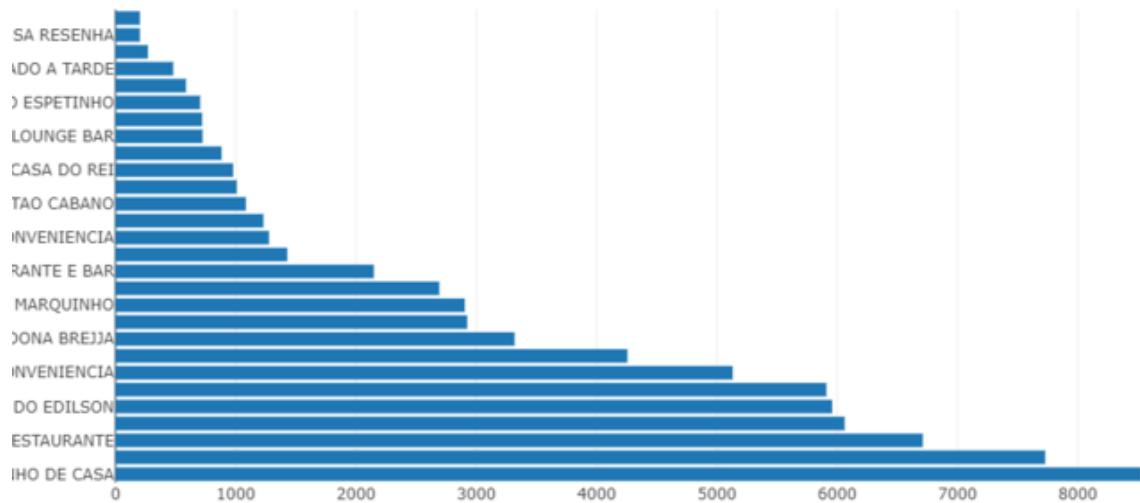
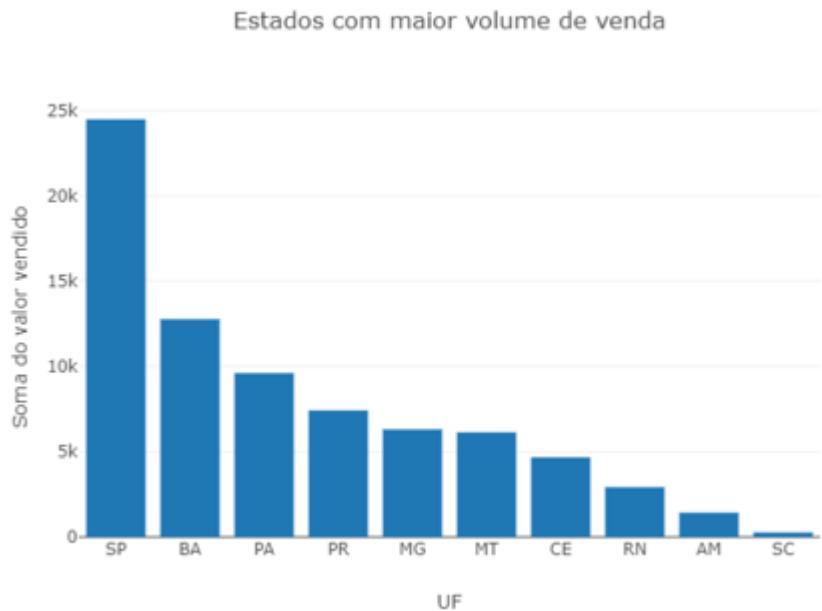
Column name	Data type	Encoding	
UF	VARCHAR	No selection	
Mapa de emp...	INTEGER	No selection	
população	INTEGER	No selection	
Relação emp...	VARCHAR	No selection	

Assim, sendo possível garantir que as tabelas sejam criadas no formato e organização desejada.

Criação de Views:

Além do upload dos dados para tabelas nos dados data warehouse também foram criadas views com base nesses dados, com intuito de trazer uma nova visão sobre eles. As views foram criadas a partir da execução de queries SQL que buscam e processam os dados presentes em nossas tabelas no data warehouse, sendo possível juntar e processar múltiplas tabelas.

Segue um exemplo das informações que foram possíveis obter através das views criadas:



Sendo assim possível, como exemplificado, obter insights como as regiões em que ocorrem a venda de certos produtos e os estabelecimentos que mais vendem.

9.6 Segurança dos dados:

O Redshift é um serviço de data warehouse que é totalmente gerenciado na nuvem da AWS, localização essa que pode se tornar perigosa, uma vez que fica vulnerável para acessos externos. Por essa razão, há uma grande ênfase na segurança dos dados, para que possam garantir a segurança dos dados guardados em nossa solução. Seguem alguns dos aspectos e ferramentas que destacam a segurança de dados oferecida pelo AWS Redshift.

1. **IAM:** O AWS IAM (identity and Access Manager) foi uma das ferramentas utilizadas para garantir a segurança dos dados presentes do Redshift, com ela é possível garantir permissões de controle e acesso aos usuários da AWS, controlando assim quem pode acessar os dados e fazer modificações no Redshift.
2. **Criptografia de dados:** Outra medida adotada para garantir a segurança dos dados é a criptografia desses dados para garantir sua integridade. Nesse processo é realizado o tipo de criptografia SSL (Secure Sockets Layer) na comunicação entre o S3 e o Redshift.
3. **Backup dos dados:** Permite a criação de screenshots automáticas do cluster Redshift, que são cópias dos dados em um ponto específico no tempo que funciona para fins de recuperação. Além disso, oferece também a capacidade de fazer backups manuais e criar cópias das screenshots para maior redundância.
4. **Logs:** É capaz de gerar logs que registram informações sobre operações realizadas no cluster, incluindo também consultas e modificações na estrutura do banco de dados, garantindo o controle sobre o que está acontecendo na aplicação.

9.7 Monitoramento do cubo de dados(CloudWatch):

O nosso objetivo é configurar o AWS CloudWatch para que ele possa monitorar o nosso cubo de dados criado no AWS Redshift para garantir o desempenho otimizado, a disponibilidade contínua e a capacidade de resposta rápida a eventos críticos. Para garantir que essas métricas sejam cumpridas, certas ações devem ser tomadas.

1. Configurar Métricas do Redshift no CloudWatch:
 - Habilitar a coleta de métricas no console do Redshift.
 - Vincular o cluster Redshift ao CloudWatch para enviar métricas automaticamente.
2. Definir alarmes no CloudWatch:
 - Configurar alarmes no CloudWatch para métricas críticas, como CPU, utilização de espaço, número de conexões, etc.
3. Logs e Análise de Desempenho:
 - Configurar a captura de logs detalhados no Redshift e enviá-los para o CloudWatch.
 - Alguns exemplos de logs que podem ser utilizados são:
 - Log de consulta (query):
 - Pode incluir detalhes como a duração da consulta, o número de linhas afetadas e o plano de execução da consulta.
 - Log de conexão (connection):
 - Pode incluir detalhes sobre quem está se conectando, de onde estão se conectando e quando as conexões foram estabelecidas e encerradas.
 - Log de erros (error):
 - Contém informações sobre erros encontrados durante a execução de consultas ou operações no cluster Redshift.
 - Log de desempenho (performance):
 - Pode incluir informações sobre o desempenho do sistema, como a utilização de recursos (CPU, memória, E/S) e tempos de resposta.

10. Análise de impacto ético

10.1 Análise de Impacto Social:

A incorporação de responsabilidade social em projetos empresariais é crucial para assegurar que as atividades comerciais não apenas alcancem metas econômicas, mas também contribuam positivamente para a sociedade e o meio ambiente. No caso do nosso projeto, que visa desenvolver um pipeline de Big Data para um distribuidor atuante em diversos setores, é fundamental considerar os impactos sociais envolvidos.

1. Positivos:

- **Eficiência Operacional:** A implementação do pipeline de Big Data pode levar a uma maior eficiência operacional para a empresa, permitindo a análise precisa do potencial de consumo em diferentes categorias e locais.
- **Geração de Empregos e Parcerias:** O desenvolvimento do projeto pode gerar oportunidades de emprego, especialmente se envolver a colaboração com instituições de ensino, como mencionado com os estudantes do Inteli.
- **Tomada de Decisões Informada:** O acesso a dados mais detalhados pode resultar em decisões mais informadas, o que, por sua vez, pode beneficiar clientes e parceiros comerciais.

2. Negativos:

- **Privacidade e Segurança dos Dados:** O manuseio de grandes volumes de dados requer atenção especial à privacidade e segurança. É necessário garantir que informações sensíveis estejam devidamente protegidas.
- **Desigualdade de Acesso:** Se o acesso ao pipeline de Big Data não for equitativo, pode haver desigualdade entre diferentes atores do setor, exacerbando disparidades já existentes.

Relação com Objetivos de Desenvolvimento Sustentável (ODS):

O projeto pode contribuir para vários Objetivos de Desenvolvimento Sustentável (ODS), como:

- **ODS 8 - Trabalho Decente e Crescimento Econômico:** A geração de empregos e a eficiência operacional podem contribuir para o crescimento econômico sustentável.
- **ODS 9 - Indústria, Inovação e Infraestrutura:** A implementação de um pipeline de Big Data representa uma inovação na infraestrutura tecnológica.
- **ODS 11 - Cidades e Comunidades Sustentáveis:** A análise granular por cidade alinha-se à meta de construir cidades mais sustentáveis.

Conclusão:

Ao incorporar a responsabilidade social no desenvolvimento do pipeline de Big Data, o projeto não apenas atenderá às necessidades do cliente, mas também proporcionará benefícios tangíveis à sociedade e ao meio ambiente. A conscientização sobre os potenciais impactos positivos e negativos e a busca contínua por práticas sustentáveis são fundamentais para assegurar que o projeto esteja alinhado com os princípios da responsabilidade social corporativa.

10.2 Viés e Discriminação:

O projeto proposto para o desenvolvimento de um pipeline de Big Data com foco em análises estatísticas para um distribuidor alimentar apresenta oportunidades significativas, mas também

implica desafios relacionados a possíveis viéses algorítmicos e discriminação involuntária. É crucial considerar esses aspectos para garantir que a implementação do sistema não prejudique grupos específicos ou exclua certas categorias, canais ou regiões.

Riscos de Viés Algorítmico:

Algoritmos de análise de dados podem incorporar viés se forem treinados com conjuntos de dados que refletem desigualdades existentes. Por exemplo, se o histórico de vendas ou comportamento de consumo usado para treinar o modelo for enviesado em relação a certos grupos demográficos, o algoritmo pode perpetuar essas discrepâncias.

Discriminação e Exclusão Involuntária:

Ao analisar dados específicos de categorias, canais e regiões, há o risco de que certos grupos ou áreas sejam inadvertidamente excluídos das análises. Isso pode resultar em decisões que não representam adequadamente a diversidade do mercado ou que prejudicam involuntariamente alguns participantes.

Estratégias para Mitigar Viés e Discriminação:

1. Diversidade nos Dados de Treinamento: Garantir que os conjuntos de dados usados para treinar o modelo incluam uma representação diversificada de todas as categorias, canais e regiões relevantes. Isso ajudará a minimizar o viés algorítmico.

2. Auditoria Regular do Modelo: Implementar procedimentos regulares de auditoria para avaliar o desempenho do modelo em relação à equidade e inclusão. Identificar e corrigir possíveis viéses que surgirem durante o uso do sistema.

3. Transparência e Interpretabilidade: Tornar o processo decisório do algoritmo transparente e comprehensível. Isso permitirá que os usuários entendam como as decisões são tomadas e identifiquem possíveis fontes de viés.

4. Envolvimento de Stakeholders Diversificados: Incluir stakeholders diversos, como representantes de diferentes regiões e segmentos de consumidores, no processo de desenvolvimento e validação do sistema. Isso proporcionará insights valiosos sobre a adequação e justiça das análises.

Objetivo de Desenvolvimento Responsável:

Além do objetivo principal de criar um pipeline eficiente para análises estatísticas, é crucial incluir o desenvolvimento responsável como parte integrante do projeto. Isso não apenas protegerá contra riscos éticos, mas também fortalecerá a reputação do distribuidor como uma empresa socialmente responsável.

Conclusão:

Ao abordar proativamente as questões de viés e discriminação no desenvolvimento do pipeline de Big Data, o projeto pode proporcionar não apenas benefícios operacionais para o distribuidor, mas também contribuir para práticas de negócios éticas e socialmente responsáveis.

10.3 Privacidade:

No âmbito do projeto que visa desenvolver um pipeline de Big Data para análises estatísticas na indústria de distribuição alimentar, é imprescindível abordar a temática da privacidade e proteção de dados. Em um cenário empresarial cada vez mais orientado por informações, a coleta,

armazenamento e uso de dados pessoais exigem uma atenção cuidadosa para garantir conformidade com regulamentações de privacidade e estabelecer práticas éticas. Aqui serão listadas importâncias de uma abordagem responsável acerca da privacidade de dados, reconhecendo que a confiança do cliente e a conformidade legal são pilares fundamentais para o sucesso e a replicabilidade do projeto em diferentes contextos empresariais.

Coleta de Dados:

O primeiro passo para garantir a privacidade é entender as formas de coleta de dados e, neste projeto, é essencial identificar quais tipos de dados serão coletados. No contexto de distribuição alimentar, envolve históricos de vendas, informações sobre a população, sua renda, entre outros. É crucial anonimizar quaisquer dados que possam ser considerados pessoais, como informações de contato, preferências de compra ou comportamentos individuais.

Armazenamento de Dados:

O armazenamento de dados também desempenha um papel vital na proteção da privacidade. Ao utilizar um data lake ou um data warehouse na AWS, é essencial implementar medidas de segurança robustas. Isso inclui o uso de criptografia, controle de acesso rigoroso e auditorias periódicas para garantir que os dados armazenados estejam protegidos contra acessos não autorizados. Vale lembrar que informações pessoais não anonimizadas devem ser armazenadas on premise, visto que dados sensíveis não devem ser compartilhados em nuvem para garantir seu sigilo.

Uso de Dados:

A análise estatística proposta no projeto, através do cubo OLAP, deve revelar insights valiosos, mas é crucial garantir que essas análises não comprometam a privacidade dos indivíduos. Deve-se remover informações identificáveis e garantir que os resultados agregados não possam ser rastreados até indivíduos específicos, no caso do cubo OLAP foram usados dados públicos, então eles atendem as exigências de privacidade.

Conformidade com Regulamentações:

O projeto deve aderir a regulamentações relevantes de privacidade, como a Lei Geral de Proteção de Dados (LGPD). Isso implica informar os participantes sobre a coleta de dados, obter seu consentimento quando necessário, e garantir que todos os processos estejam em conformidade com as diretrizes estabelecidas pelas autoridades de proteção de dados.

Conclusões:

Logo, integrar considerações de privacidade e proteção de dados desde o início do projeto não apenas atende a requisitos regulatórios, mas também estabelece uma base sólida para o sucesso e a aceitação do cliente. Essa abordagem ética contribui para a construção de relacionamentos duradouros e para a replicabilidade da solução em diferentes cenários de negócios.

10.4 Equidade e justiça:

No contexto do projeto, é imperativo dedicar uma atenção especial à equidade e à justiça. Ao lidar com dados que influenciam diretamente as decisões operacionais e táticas do cliente, a abordagem deve transcender a eficácia pura e abranger a responsabilidade ética. Examina-se, portanto, a necessidade de garantir que o pipeline não apenas otimize as ações comerciais, mas também minimize disparidades, promovendo a equidade e a justiça em todas as fases do processo.

Possíveis Impactos em Grupos Específicos:

- **Disparidades Regionais:** Diferentes regiões podem apresentar características socioeconômicas distintas, influenciando os padrões de consumo. O pipeline de Big Data, se não considerar essas diferenças, pode resultar em estratégias desproporcionais, prejudicando ou favorecendo certas regiões em detrimento de outras.
- **Viés Socioeconômico:** A POF pode ter um viés socioeconômico, pois as famílias mais vulneráveis podem ter menor representatividade na pesquisa. Se o pipeline basear suas análises apenas nesses dados, pode subestimar as necessidades e padrões de consumo de grupos de menor poder aquisitivo.
- **Diferenças Regionais na Amostra:** A amostra da POF pode não ser uniformemente distribuída geograficamente. Isso pode levar a uma falta de representação de certas regiões, impactando a precisão das análises para áreas específicas do país.

Minimização de Disparidades:

- **Segmentação Precisa dos Dados:** Ao coletar dados, é crucial segmentar de forma precisa, considerando variáveis como localização, tamanho do estabelecimento, perfil socioeconômico da região, entre outros. Isso assegura que as análises refletem a diversidade dos contextos e evite generalizações inadequadas.
- **Envolvimento Direto de Stakeholders Representativos:** Garantir a participação direta de stakeholders representativos de diferentes grupos na concepção e revisão do pipeline é fundamental. Isso ajuda a validar as decisões, considerando perspectivas diversas e garantindo que as estratégias sejam justas para todos.
- **Métricas de Avaliação da Equidade:** Implementar métricas específicas para avaliar a equidade nas decisões e resultados do pipeline, ajustando continuamente com base nessas métricas, promove a adaptação constante e a correção de possíveis disparidades.
- **Ponderação Adeuada dos Dados:** Ao utilizar dados da POF, é crucial aplicar técnicas de ponderação que compensam desequilíbrios na amostra. Isso ajuda a garantir que as análises refletem com mais precisão as características demográficas e econômicas da população em diferentes regiões.

10.5 Transparência e Consentimento:

A transparência e o consentimento informado são princípios fundamentais em qualquer projeto, especialmente quando se trata de lidar com dados sensíveis e estratégicos, como no caso do pipeline de Big Data proposto para o distribuidor alimentar. A garantia de que todas as partes envolvidas têm acesso claro às informações relevantes e que o consentimento é obtido de maneira adequada é crucial para o sucesso e a ética do projeto.

Transparência:

No contexto desse projeto, a transparência se refere à clareza e acessibilidade das informações relacionadas à coleta, processamento e uso dos dados. Para assegurar a transparência, é essencial que o cliente distribuidor seja plenamente informado sobre o escopo do projeto, os tipos de dados coletados, os métodos de processamento e análise estatística empregados e os resultados esperados. Isso permite que o cliente compreenda completamente como as informações serão utilizadas para atender sua demanda pelo desenvolvimento de um método de trabalho para implantação do Pipeline de Big Data.

Consentimento Informado:

O consentimento informado é a garantia de que o cliente está ciente e concorda com a coleta e o uso dos dados conforme proposto pelo projeto. No contexto deste pipeline de Big Data, o cliente distribuidor deve ser informado sobre a finalidade específica da análise estatística, os potenciais benefícios que serão obtidos e quaisquer implicações associadas à divulgação dessas informações.

Conclusões:

Logo, a responsabilidade de garantir a transparência recai à coleta e ao uso dos dados, enquanto o consentimento de dados é garantido no momento do armazenamento. Dessa forma, o uso de dados públicos garante o consentimento. Além disso, ao construir o infográfico, será necessário garantir que a visualização dos dados não possibilite análises ambíguas.

11. Conclusões

12. Referências

13. Anexos