



BIG DATA

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA

INTEGRATION

Autores:

Gabriela de Moraes da Silva

Larissa Carvalho

Sophia Dias

Thainá Lima

Ueliton Rocha

Vitória Rodrigues de Oliveira

Data de criação: 24 de Outubro de 2023

SÃO PAULO – SP

2023

Sumário

Sumário	3
1. Introdução	9
1.1. Parceiro de Negócios	9
1.2. Definição do Problema	9
1.2.1. Problema	9
2. Objetivos	9
2.1. Objetivos Gerais	10
2.2. Objetivos Específicos	10
2.3. Justificativa	10
3. Compreensão do Problema	10
3.1. Proposta de Valor	11
3.1.1. Perfil do Cliente	11
Tarefas do cliente:	11
Dores:	12
Ganhos:	13
3.1.2. Proposta de Valor	13
Produtos e Serviços:	13
Aliviam as dores:	13
Criadores de ganho:	14
3.2 Matriz de Risco	15
Sprint 1:	15
Sprint 2:	16
Sprint 3:	17
Sprint 4:	18
Sprint 5:	19
3.3 TAM SAM SOM	21
4. Análise de Experiência do Usuário	23
4.1 Personas	23
4.1.1. Persona 1: Juliana Santos	23
Dores:	24
Necessidades:	25
Nível de letramento digital:	25
Exemplos de uso e preferências:	25
Citações a partir dos pensamentos de Juliana:	26
4.1.2. Persona 2: Augusto Ribeiro	27
Dores:	28
Necessidades:	29
Nível de letramento digital:	29

Cenários de Interação:	29
Exemplos de uso e preferências:	30
Citações a partir dos pensamentos de Augusto:	30
4.2 Jornada do Usuário	30
4.2.1. Jornada do usuário persona 1 - Juliana Santos, Consultora de Marketing e Vendas.	31
4.2.2. Jornada do usuário persona 2 - Augusto Ribeiro, Tech Digital	32
4.3 User Stories	33
User story 1	33
Descrição User Story 1	34
User story 2	35
Descrição User Story 2	36
User story 3	37
Descrição User Story 3	38
User story 4	39
Descrição User Story 4	40
User story 5	41
Descrição User Story 5	42
User story 6	43
Descrição User story 6	44
User story 7	45
Descrição User story 7	46
User story 8	47
Descrição User Story 8	48
User story 9	49
Descrição User Story 9	50
User story 10	51
Descrição User Story 10	52
User story 11	53
Descrição User Story 11	54
User story 12	55
Descrição User Story 12	56
4.4.Wireframe	57
5. Arquitetura Macro	59
5.1 Componentes da Arquitetura	61
5.1.1.API do cliente	61
5.1.2.CSVs	61
5.1.3.Script Python	61
5.1.4.Bucket S3	61
5.1.5.Lambda	61
5.1.6.Spark	61
5.1.7.VPC (Virtual Private Cloud)	62

5.1.8.AWS CloudWatch	62
5.1.9.AWS Redshift	62
5.1.10.Segurança	63
5.2.Expurgo dos dados	63
5.2.Análise Descritiva	64
5.2.1.Base dos dados.org - Base dos Dados	65
5.2.2.IBGE Dados Abertos - Dados Abertos IBGE	65
5.2.3.POF (pesquisa orçamento familiar - POF 2017-2018 IBGE	65
5.2.3.1.Tabelas POF	65
5.2.4.RAIS e CAGED Microdados - Microdados RAIS e CAGED — Ministério do Trabalho e Emprego (www.gov.br)	68
5.2.5.Receita Federal Dados Abertos - Dados Abertos — Receita Federal (www.gov.br)	68
5.2.6.Dados Abertos - MEC - Página inicial (www.gov.br)	68
5.2.7.Microdados — Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira Inep (www.gov.br)	69
5.2.8.Bem vindo - OPENDATASUS (saude.gov.br)	69
5.2.8.Zip and Postal Codes of All Countries Database Hub (back4app.com)	69
5.2.9.API - JSON - BASE DE VENDAS FICTÍCIA VIA ENDPOINT INTEGRATION	
70	
5.3.Análise Exploratória	70
6. Desenvolvimento do Cubo de Dados	71
6.1. Processo de ETL para Cubo de Dados OLAP	71
6.1.1.Visão geral	71
6.2.Processo de ETL	71
6.2.1.Extração de dados	71
6.2.2.Transformação de dados	72
6.2.3.Carregamento no AWS S3	72
6.3.Armazenamento e análise de dados	73
6.3.1.Configuração do AWS Redshift	76
6.3.2.Estrutura do Data Warehouse: Arquitetura de Três Fases	81
6.3.2.1.Utilização de Views no Data Warehouse	81
6.4.Aspectos de segurança, privacidade e conformidade	82
6.5.Monitoramento e gerenciamento do processo de ETL	83
6.6.VIEWS	83
6.6.1. Consumo Individual	84
6.6.2. Características do Consumo	84
6.6.3. CNPJ s e Sales	84
6.6.4.CNPJs, Sales e Category	84
7. Curadoria dos dados	85
7.1. POF (Pesquisa de Orçamento Familiar)	85
7.2. CNPJ	85

7.3. CEP	85
8. Modelos Ensemble	86
8.1 Algoritmo K-means	87
8.2 Modelo Random Forest	90
8.3 Métodos de avaliação e validação com o CRISP-DM	91
8.4 Proposta de otimização do modelo	92
9. Infográficos	93
9.1 Objetivos	93
9.2 Gráficos iniciais com o Power BI	94
9.3 Gráficos gerados no Metabase	95
9.4 Infográficos Criados no Metabase	95
9.5 Relatório de Análise de Eficácia dos Infográficos	96
9.5.1.Sugestões de Melhoria:	97
Gráfico 1 - Mapa de maior venda em cada estado	97
Gráfico 2 - Venda de carne por tempo	98
Gráfico 3 - Média de despesa por horário de consumo	98
Gráfico 4 - Refeições mais realizadas	99
Gráfico 5 - Categoria mais vendida	99
Gráfico 6 - Quantidade consumida por hora	100
Gráfico 7 - Consumo por local de refeição em dias atípicos	100
Gráfico 8 - Média de valor de compra por estado	101
Gráfico 9 - Tipo de dieta por estado	102
9.5.2.Conclusão	102
10. Análise financeira	103
10.1.Receita	103
10.2.Orçamento total	103
10.3.Custos	103
10.3.1.Métodos de pagamento AWS	103
10.3.2.Custo dos Serviços utilizados	104
10.3.3.Mapeamento de custos 12 meses AWS:	105
10.3.4.Análise comparativa entre Amazon Web Services (AWS) e Microsoft Azure	105
10.3.5.Custos de desenvolvimento	108
11. Impacto Ético	110
11.1. Introdução	110
11.2. Privacidade e Proteção de Dados	110
11.3. Equidade e justiça	111
11.4.Transparência em Projetos de Big Data	111
11.5. Estratégias para Garantir Transparência e Consentimento em Big Data	112
11.6.Responsabilidade social	113
11.7. Tratamento Ético de Dados Sensíveis	114
11.8. Viés Algorítmico e Discriminação em Projetos de Big Data	114

11.9. Estratégias para Mitigar Viés e Discriminação	114
11.10. Considerações Finais	115
12. Plano de Comunicação	115
12.1 Objetivo	115
12.2 Stakeholders	117
12.3 Mensagens Chave	117
12.4 Canais de Comunicação:	117
12.5 Plano de Implementação	118
12.6 Medidas de Sucesso	119
12.7 Feedback e Ajustes	119
13. Conclusões	122
14. Referências	123
15. Anexos	124

1. Introdução

A Integration, consultoria global de estratégia e gestão empresarial, destacou um grande desafio estratégico. Este desafio centra-se na necessidade urgente de uma ferramenta que permita que os consultores de marketing/vendas compreendam detalhadamente o potencial de consumo dos produtos dos seus clientes, filtrando por categorias, canais e geografias.

1.1. Parceiro de Negócios

O parceiro do projeto é a Integration, uma consultoria estratégica localizada em São Paulo. A empresa desempenha uma entrega de projetos tanto no Brasil quanto em outras regiões, incluindo América Latina, Europa e Estados Unidos. Seu foco abrange diversas áreas, como Marketing & Vendas, Finanças & Gestão, Cadeia de Suprimentos, Sustentabilidade e Implementação de Estratégias. A empresa tem uma forte presença em setores como Bens de Consumo, Varejo, Private Equity & Investimentos, Financeiro & Pagamentos, Indústria, Agronegócio e Farmacêutico/Saúde.

1.2. Definição do Problema

A capacidade dos consultores da Integration em realizar análises estratégicas direcionadas é comprometida pela ausência de informações cruciais sobre categorias, canais e regiões. Essa lacuna prejudica a formulação de estratégias eficazes, impedindo uma compreensão abrangente do cenário de consumo.

1.2.1. Problema

O problema central identificado é a falta de uma ferramenta que possibilite a compreensão granular do potencial de consumo de produtos, prejudicando a capacidade dos consultores em conduzir análises estratégicas direcionadas.

2. Objetivos

O projeto tem como objetivo principal a criação de um pipeline de Big Data na AWS para realizar análises estatísticas em dados armazenados em um datalake ou data warehouse. Além disso, busca-se desenvolver um infográfico que permitirá aos consultores da Integration tomar decisões mais informadas no seu dia-a-dia.

2.1.Objetivos Gerais

1. Desenvolver um pipeline de Big Data na AWS;
2. Desenvolvimento de um cubo de dados;
3. Criar um infográfico para os consultores de marketing/vendas da Integration.

2.2.Objetivos Específicos

1. Consolidar dados de consumo categorizados;
2. Criar um infográfico informativo e fácil de entender;
3. Incorporar informações geográficas para uma análise regional do consumo;
4. Desenvolver uma interface intuitiva para visualização dos resultados;
5. Integrar detalhes sobre canais de distribuição nos conjuntos de dados;
6. Implementar algoritmos estatísticos para a análise dos dados.

2.3.Justificativa

O projeto tem como objetivo não apenas preencher as lacunas identificadas na capacidade de criar estratégias dos consultores, mas também aprimorar suas análises e decisões cotidianas. A integração dessas informações detalhadas resultará em uma visão abrangente do cenário de consumo, permitindo que a Integration forneça soluções estratégicas mais precisas e personalizadas para seus clientes. Isso contribuirá para consolidar ainda mais sua posição como consultoria de destaque no setor. Além disso, o projeto simplifica o dia-a-dia da equipe de tech and digital. Isso ocorre ao evitar retrabalho e processamento manual de dados, uma vez que o projeto envolve a implementação de uma ferramenta padronizada que automatiza significativamente esse processo.

3.Compreensão do Problema

Esta seção é crucial para fundamentar a abordagem estratégica. Na "Proposta de Valor", destacamos singularidades que conferem valor aos clientes. A "Matriz de Risco" antecipa desafios, permitindo estratégias de mitigação, enquanto a análise "TAM SAM SOM" define o mercado total, potencial alcançável e objetivos realizáveis. Esses elementos fornecem a base para decisões e ações estratégicas, delineando claramente nosso caminho para soluções eficazes e adaptáveis às necessidades do público-alvo.

3.1. Proposta de Valor

O Canvas Proposta de Valor é uma ferramenta de gestão estratégica que auxilia na criação e desenvolvimento de novos produtos, serviços e negócios. Essa ferramenta é dividida em dois blocos: o perfil do cliente e a proposta de valor.

No bloco de proposta de valor, é necessário definir qual é o produto desenvolvido, quais são os principais criadores de ganho que o cliente tem ao adquirir o produto e quais são os aliviadores de dores que o cliente terá. Já no perfil do cliente, deve-se identificar quais são as tarefas que ele irá realizar, quais são os principais ganhos e dores do atual processo. Abaixo segue o Canvas Proposta de Valor completo.

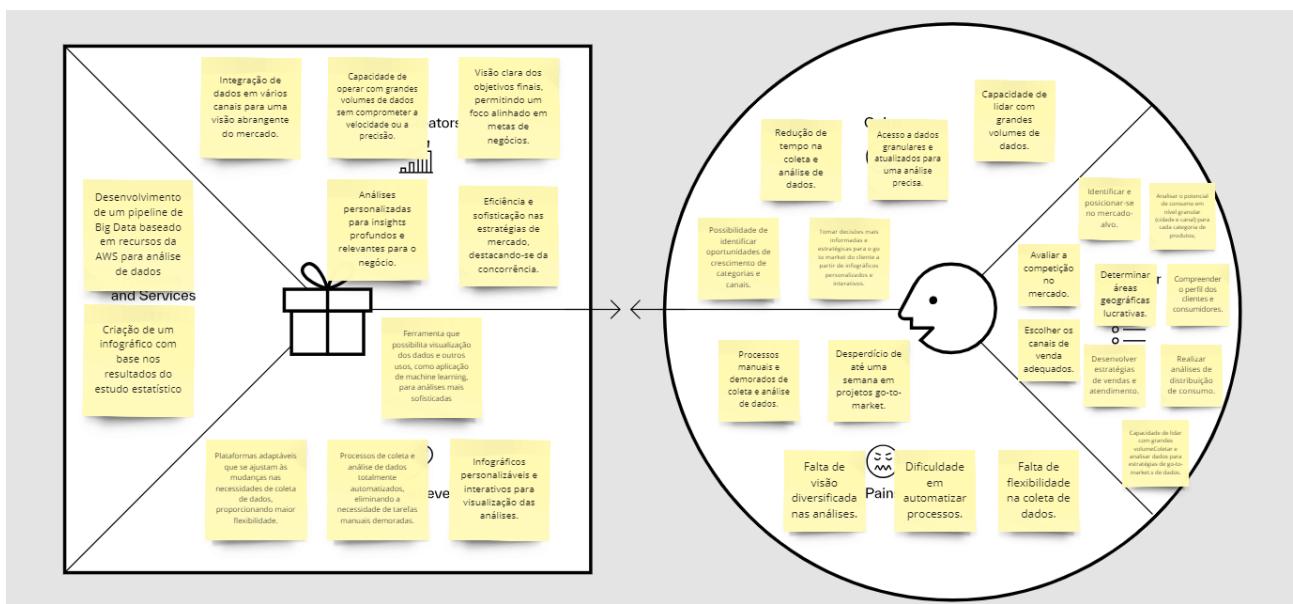


Imagen 1 - Canva da proposta de valor

3.1.1. Perfil do Cliente

Tarefas do cliente:

- Analisar o potencial de consumo em nível granular (cidade e canal) para cada categoria de produtos: Isso significa que o cliente precisa avaliar o potencial de consumo para categorias específicas de produtos em diferentes cidades e canais de venda.
 - Identificar e posicionar-se no mercado-alvo: O cliente deseja determinar qual segmento de mercado é mais adequado para seus produtos e como se posicionar efetivamente nesse mercado.

- Avaliar a competição no mercado: Compreender a concorrência é essencial para se diferenciar e oferecer algo único ao mercado.
- Escolher os canais de venda adequados: Dependendo do produto e do público-alvo, diferentes canais de venda podem ser mais eficazes.
- Determinar áreas geográficas lucrativas: O cliente deseja focar suas estratégias em regiões que ofereçam o maior retorno sobre o investimento.
- Compreender o perfil dos clientes e consumidores: Isso envolve entender as necessidades, desejos e comportamentos dos consumidores para atendê-los melhor.
- Desenvolver estratégias de vendas e atendimento: Com base nas informações coletadas, o cliente precisa elaborar planos de ação eficazes para vendas e atendimento ao cliente.
- Realizar análises de distribuição de consumo: Entender como o consumo está distribuído entre diferentes segmentos e regiões pode ajudar a direcionar esforços de marketing e vendas.
- Coletar e analisar dados para estratégias de go-to-market: Antes de lançar um produto ou entrar em um novo mercado, é crucial coletar e analisar dados relevantes para informar a estratégia.

Dores:

Estas são as dificuldades e desafios que o cliente enfrenta em suas operações diárias:

- Processos manuais e demorados de coleta e análise de dados: Sem automação, a coleta e análise de dados podem ser extremamente demoradas e propensas a erros.
- Desperdício de até uma semana em projetos go-to-market: O tempo é um recurso valioso, e qualquer atraso pode resultar em oportunidades perdidas.
- Falta de visão diversificada nas análises: Se os dados não forem abrangentes ou representativos, as análises podem não fornecer insights completos ou precisos.
- Dificuldade em automatizar processos: A falta de ferramentas ou conhecimento para automatizar tarefas pode impedir a eficiência operacional.
- Falta de flexibilidade na coleta de dados: Sem sistemas adaptáveis, pode ser difícil ajustar os métodos de coleta de dados à medida que as necessidades mudam.

Ganhos:

Estes são os benefícios e vantagens que o cliente espera obter ao resolver suas "dores":

- Capacidade de lidar com grandes volumes de dados: Isso permite ao cliente processar e analisar grandes conjuntos de dados sem comprometer a velocidade ou precisão.
- Acesso a dados granulares e atualizados para uma análise precisa: Dados detalhados e atualizados podem fornecer insights mais precisos e valiosos.
- Redução de tempo na coleta e análise de dados: A eficiência é crucial para permanecer competitivo e responder rapidamente às mudanças do mercado.
- Possibilidade de identificar oportunidades de crescimento de categorias e canais: Ao entender onde estão as oportunidades, o cliente pode direcionar seus recursos de forma mais eficaz.
- Tomar decisões mais informadas e estratégicas para o go to market do cliente a partir de infográficos personalizados e interativos: Visualizações claras e personalizadas podem ajudar a informar decisões estratégicas e proporcionar uma vantagem competitiva.

3.1.2. Proposta de Valor

Produtos e Serviços:

- Desenvolvimento de um pipeline de Big Data baseado em recursos da AWS para análise de dados: Isso sugere a criação de uma infraestrutura robusta e escalável na AWS para processar e analisar grandes volumes de dados.
- Criação de um infográfico com base nos resultados do estudo estatístico: A visualização de dados através de infográficos pode ajudar a interpretar e comunicar resultados complexos de forma mais intuitiva.

Aliviam as dores:

Estas soluções são projetadas especificamente para abordar e aliviar as "dores" mencionadas anteriormente:

- Ferramenta que possibilita visualização dos dados e outros usos, como aplicação de machine learning, para análises mais sofisticadas: Isso sugere uma plataforma

multifuncional que não apenas visualiza, mas também aplica técnicas avançadas de análise.

- Infográficos personalizáveis e interativos para visualização das análises: Oferecer visualizações customizáveis permite que os clientes vejam os dados da maneira que faz mais sentido para eles.
- Processos de coleta e análise de dados totalmente automatizados, eliminando a necessidade de tarefas manuais demoradas: Automatizar esses processos pode economizar tempo e recursos significativos.
- Plataformas adaptáveis que se ajustam às mudanças nas necessidades de coleta de dados, proporcionando maior flexibilidade: Ter uma solução flexível permite ao cliente adaptar-se rapidamente às mudanças nas condições de mercado ou requisitos de negócios.

Criadores de ganho:

Estes são os benefícios adicionais que a proposta oferece, que não apenas aliviam as dores, mas também oferecem vantagens adicionais:

- Eficiência e sofisticação nas estratégias de mercado, destacando-se da concorrência: Isso sugere uma vantagem competitiva clara através do uso de análises avançadas.
- Análises personalizadas para insights profundos e relevantes para o negócio: Ao oferecer análises personalizadas, os clientes podem obter insights que são especialmente relevantes para seus negócios específicos.
- Visão clara dos objetivos finais, permitindo um foco alinhado em metas de negócios: Ter uma direção clara pode ajudar a alinhar toda a organização em direção a metas comuns.
- Capacidade de operar com grandes volumes de dados sem comprometer a velocidade ou a precisão: Isso sugere uma solução escalável que pode crescer com as necessidades do cliente.
- Integração de dados em vários canais para uma visão abrangente do mercado: Ao integrar dados de várias fontes, o cliente pode obter uma visão mais completa do mercado e de seus clientes.

3.2 Matriz de Risco

É uma das principais ferramentas na análise de negócios, utilizada para identificar, avaliar e priorizar os riscos operacionais existentes na solução desenvolvida com base em sua probabilidade e impacto. Ela contribui para a tomada de decisões mais informadas e eficazes no projeto pela equipe de desenvolvimento, pois facilita a priorização de riscos e o desenvolvimento de estratégias de mitigação ou prevenção.

Enquanto a matriz de risco oferece uma visão das possíveis contingências e obstáculos que podem surgir, o plano de ação representa a resposta estratégica para enfrentar esses desafios.

A figura abaixo ilustra a construção da matriz de risco e seu plano de ação para o projeto.

Sprint 1:

Matriz de Risco											
Probabilidade	Riscos						Oportunidades				
	Muito Alta	1				Custos elevados referente a arquitetura em cloud	Capacidade insuficiente do computador para processar a quantidade de dados				
	Alta	2						Melhoria da eficiência operacional da empresa economizando recursos	Redução dos custos atuais do cliente com ferramentas cloud.		
	Média	3			Pouca usabilidade do infográfico pelos usuários massivos de dados	Ausência dos integrantes no dev	Baixo conhecimento técnico entre grande parte os membros da equipe na arquitetura do projeto		Desenvolvimento de um infográfico com interface intuitiva, facilitando a tomada de decisões.		
	Baixa	4			Integração dos dados demandar muito tempo e atrasar entrega do projeto	Dificuldade na execução de todas as tarefas durante a sprint mais curta.	Baixa qualidade dos dados levando a resultados imprecisos e incorretos		Optimização do processamento de grandes volumes de dados permitindo re-trabalho		
	Muito Baixa	5			Dependência de aulas dos professores para dar continuidade ao projeto	Exposição de dados sensíveis pelo github					
		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											

Imagen 2 - Matriz de risco 1

Matriz	Descrição do Impacto	Responsável	Descrição da Ação
1-5	A incapacidade do computador em lidar com a grande quantidade de dados pode resultar em atrasos no pré-processamento, impactando a eficiência da análise e o cronograma do projeto.	Thainá	Garantir o pleno funcionamento da máquina virtual compatível com a utilização para que recursos sejam adequados para o volume de dados do projeto.
3-3	A dificuldade em tornar o infográfico compreensível pode comprometer a utilidade do projeto, dificultando a análise de informações valiosas.	Vitória	Estar sempre em contato com o professor de UX para obter feedback sobre a usabilidade e com o cliente para saber se suas dores foram solucionadas
3-4	Faltas dos integrantes não avisadas antecipadamente podem acarretar no atraso de tarefas e sobrecarga de outros integrantes	Larissa	Comunicação clara e notificação antecipada em caso de ausência. Certificar que todos os membros da equipe compreendem a necessidade de informar sobre ausências.
3-5	Dificuldades técnicas dos integrantes sem aviso prévio pode atrasar o andamento e entrega do projeto além de prejudicar o engajamento do grupo	Gabriela	Garantir que os POs do time estejam a par do problema, a fim de que o professor de programação seja recorrido para auxiliar o time. Além disso, os autoestudos devem ser revisitados em primeira instância.
1-4	A quantidade de espaço na nuvem pode aumentar os custos do projeto, impactando o orçamento e a viabilidade financeira.	Larissa	Identificar e eliminar recursos subutilizados ou desnecessários nas clouds, como instâncias ociosas, volumes de armazenamento não utilizados e bancos de dados não críticos.
4-3	Integração dos dados pode demandar muito tempo e resultar em atrasos significativos na entrega do projeto	Thainá	Ao perceber que o problema está acontecendo, é substancial o estudo de outras ferramentas que realizem o processamento e integração dos dados de forma mais facilitada.
4-4	Dificuldade na execução de todas as tarefas durante a sprint mais curta pode resultar em atrasos no projeto, comprometendo prazos e entregas.	Vitória	Acompanhar o progresso da sprint regularmente por meio da daily e cerimônias do Scrum.
4-5	A baixa qualidade dos dados podem afetar a credibilidade do projeto e as conclusões resultadas dos dados.	Ueliton	Aplicação de conceitos matemáticos, como o PCA que permite a ampliação da qualidade dos dados utilizados e garantir melhores resultados.
5-3	Dependência de aulas dos professores para dar continuidade ao projeto pode resultar em atrasos na progressão do projeto.	Sophia	Conduzir uma avaliação das necessidades individuais de aprendizado de cada membro da equipe e identificar se há necessidade de auto-estudos extras.
5-4	Quebra de licenças e exposição de dados sensíveis pelo github pode resultar em perda de dados e danos à reputação do projeto.	Ueliton	Retirada dos dados da internet com o máximo de urgência e anonimação imediata.

Imagen 3 - Descrição 1

Sprint 2:

Matriz de Risco											
Probabilidade	Riscos						Oportunidades				
	Muito Alta	1	Integração dos dados demandar muito tempo e atrasar entrega do projeto	Custos elevados referente a arquitetura em cloud	Capacidade insuficiente de computador para processar a quantidade de dados						
	Alta	2	Mudanças periódicas no nosso planejamento, dado a essa serem desenvolvidas	Dificuldade na execução de todas as tarefas durante a sprint mais curta	Meioria da eficiência operacional da empresa economizando recursos	Redução dos custos atuais do cliente com ferramentas cloud.	Capacitação do time em Machine Learning com foco em algoritmos relacionados a Big Data.				
	Média	3	Pouca usabilidade do infográfico pelo número massivo de dados	Ausência dos integrantes acima de 50% das atividades nos integrantes	Baixo conhecimento e habilidade das mentes da equipe na arquitetura do projeto	Desenvolvimento de um interface de usuário mais intuitiva, facilitando a tomada de decisões.					
	Baixa	4	Oportunidade de aulas dos professores para dar continuidade ao projeto	Exposição de dados sensíveis pelo github	Baixa quantidade de dados levando a resultados imprecisos e incorretos	Optimização do processamento de grandes volumes de dados para melhor desempenho					
	Muito Baixa	5									
		1	2	3	4	5	5	4	3	2	1
Impacto											

Imagen 4 - Matriz de risco 2

Plano de Ação			
Matriz	Descrição do Impacto	Responsável	Descrição da Ação
1-5	A incapacidade do computador em lidar com a grande quantidade de dados pode resultar em atrasos no pré-processamento, impactando a eficiência da análise e o cronograma do projeto.	Thainá	Garantir o pleno funcionamento da máquina virtual compatível com a utilização para que recursos sejam adequados para o volume de dados do projeto.
3-3	A dificuldade em tornar o infográfico comprehensível pode comprometer a utilidade do projeto, dificultando a análise de informações valiosas.	Vitória	Estar sempre em contato com o professor de UX para obter feedback sobre a usabilidade e com o cliente para saber se suas dores foram solucionadas
3-4	Faltas dos integrantes não avisadas antecipadamente podem acarretar no atraso de tarefas e sobrecarga de outros integrantes	Larissa	Comunicação clara e notificação antecipada em caso de ausência. Adaptação do horário do membro do time que terá faltas recorrentes para que suas atividades sejam cumpridas dentro do prazo da atividade
3-5	Dificuldades técnicas dos integrantes sem aviso prévio pode atrasar o andamento e entrega do projeto além de prejudicar o engajamento do grupo	Gabriela	Garantir que os POs do time estejam a par do problema, a fim de que o professor de programação seja recorrido para auxiliar o time. Além disso, os autoestudos devem ser revisados em primeira instância.
1-4	A quantidade de espaço na nuvem pode aumentar os custos do projeto, impactando o orçamento e a viabilidade financeira.	Larissa	Identificar e eliminar recursos subutilizados ou desnecessários nas clouds, como instâncias ociosas, volumes de armazenamento não utilizados e bancos de dados não críticos.
1-3	Integração dos dados pode demandar muito tempo e resultar em atrasos significativos na entrega do projeto	Thainá	Ao perceber que o problema está acontecendo, é substancial o estudo de outras ferramentas que realizem o processamento e integração dos dados de forma mais facilitada.
2-5	Dificuldade na execução de todas as tarefas durante a sprint mais curta pode resultar em atrasos no projeto, comprometendo prazos e entregas.	Vitória	Priorização das atividades conforme as entregas e garantir que as mesmas sejam finalizadas na data estabelecida.
4-5	A baixa qualidade dos dados podem afetar a credibilidade do projeto e as conclusões resultadas dos dados.	Ueliton	Aplicação de conceitos matemáticos, como o PCA, que permite a ampliação da qualidade dos dados utilizados e garantir melhores resultados.
5-3	Dependência de aulas dos professores para dar continuidade ao projeto pode resultar em atrasos na progressão do projeto.	Sophia	Conduzir uma avaliação das necessidades individuais de aprendizado de cada membro da equipe e identificar se há necessidade de auto-estudos extras.
5-4	Quebra de licenças e exposição de dados sensíveis pelo github pode resultar em perda de dados e danos à reputação do projeto.	Ueliton	Retirada dos dados da internet com o máximo de urgência e anonimação imediata.
2-4	Mudanças periódicas no nosso planejamento devido à quantidade de atividades pode causar estresse adicional à equipe e dificultar o cumprimento dos prazos.	Vitória	Validação de cada artefato da sprint com o professor responsável através do grupo no slack.

Imagen 5 - Descrição 2

Sprint 3:

Matriz de Risco											
Probabilidade	Riscos						Oportunidades				
	Muito Alta	1			Integração dos dados demora muito tempo e atrasar entrega do projeto	Custos elevados referente a arquitetura em cloud	Perda ou Exclusão da Conta da AWS e seus Dados Associados				
	Alta	2			Ausência dos integrantes no desacoplamento, acumulo de atividades nos integrantes	Mudanças periódicas no nosso planejamento, dado a quantidade de atividades a serem desenvolvidas	Dificuldade na execução de todas as tarefas durante a sprint mais curta				
	Média	3			Pouca usabilidade do infográfico pelo número massivo de dados	Capacidade insuficiente do computador para processar a quantidade de dados	Baixo conhecimento técnico individual entre os membros da equipe na arquitetura do projeto				
	Baixa	4					Baixa qualidade dos dados levando a resultados imprecisos e incorretos				
	Muito Baixa	5			Dependência de aulas dos professores para dar continuidade ao projeto	Exposição de dados sensíveis pelo github					
Impacto											
	1	2	3	4	5	5	4	3	2	1	
Muito Baixa	Baixa	Média	Alta	Muito Alta	Muito Alta	Muito Alta	Alta	Média	Baixa	Muito Baixa	

Imagen 6 - Matriz de risco 3

Plano de Ação				
Matriz	Descrição do Impacto		Responsável	Descrição da Ação
3-4	A incapacidade do computador em lidar com a grande quantidade de dados pode resultar em atrasos no pré-processamento, impactando a eficiência da análise e o cronograma do projeto.		Thainá	Garantir o pleno funcionamento da máquina virtual compatível com a utilização para que recursos sejam adequados para o volume de dados do projeto.
3-3	A dificuldade em tornar o infográfico compreensível pode comprometer a utilidade do projeto, dificultando a análise de informações valiosas.		Vitória	Estar sempre em contato com o professor de UX para obter feedback sobre a usabilidade e com o cliente para saber se suas dores foram solucionadas
2-3	Faltas dos integrantes não avisadas antecipadamente podem acarretar no atraso de tarefas e sobrecarga de outros integrantes		Larissa	Comunicação clara e notificação antecipada em caso de ausência. Adaptação do horário do membro do time que terá faltas recorrentes para que suas atividades sejam cumpridas dentro do prazo da atividade
3-5	Dificuldades técnicas dos integrantes sem aviso prévio pode atrasar o andamento e entrega do projeto além de prejudicar o engajamento do grupo		Gabriela	Garantir que os POs do time estejam a par do problema, a fim de que o professor de programação seja recorrido para auxiliar o time. Além disso, os autoestudos devem ser revisitados em primeira instância.
1-4	A quantidade de espaço na nuvem pode aumentar os custos do projeto, impactando o orçamento e a viabilidade financeira.		Larissa	Identificar e eliminar recursos subutilizados ou desnecessários nas clouds, como instâncias ociosas, volumes de armazenamento não utilizados e bancos de dados não críticos.
1-3	Integração dos dados pode demandar muito tempo e resultar em atrasos significativos na entrega do projeto		Thainá	Ao perceber que o problema está acontecendo, é substancial o estudo de outras ferramentas que realizem o processamento e integração dos dados de forma mais facilitada.
2-5	Dificuldade na execução de todas as tarefas durante a sprint mais curta pode resultar em atrasos no projeto, comprometendo prazos e entregas.		Vitória	Priorização das atividades conforme as entregas e garantir que as mesmas sejam finalizadas na data estabelecida.
4-5	A baixa qualidade dos dados podem afetar a credibilidade do projeto e as conclusões resultadas dos dados.		Ueliton	Aplicação de conceitos matemáticos, como o PCA, que permite a ampliação da qualidade dos dados utilizados e garantir melhores resultados.
5-3	Dependência de aulas dos professores para dar continuidade ao projeto pode resultar em atrasos na progressão do projeto.		Sophia	Conduzir uma avaliação das necessidades individuais de aprendizado de cada membro da equipe e identificar se há necessidade de auto-estudos extras.
5-4	Quebra de licenças e exposição de dados sensíveis pelo github pode resultar em perda de dados e danos à reputação do projeto.		Ueliton	Retirada dos dados da internet com o máximo de urgência e anonimação imediata.
2-4	Mudanças periódicas no nosso planejamento devido à quantidade de atividades pode causar estresse adicional à equipe e dificultar o cumprimento dos prazos.		Vitória	Validação de cada artefato da sprint com o professor responsável através do grupo no slack.
1-5	A exclusão da conta da AWS pode resultar na perda irreversível de todos os dados, trabalhos, bases e recursos associados à plataforma.		Thainá	Utilizar os labs de cada integrante e manter uma conta adicional na AWS como backup.

Oportunidades

Até o momento, não foram identificadas oportunidades.

Imagen 7 - Descrição 3

Sprint 4:

Matriz de Risco											
Probabilidade	Riscos						Oportunidades				
	Muito Alta	1	Integração dos dados demandar muito tempo e atrasar a entrega do projeto	Ausência dos integrantes que dev acarretarem atrasos na entrega dos integrantes	Perda ou Exclusão da Conta da AWS e seus Dados Associados	Não possuir uma conta ativa na AWS					
	Alta	2		pouca usabilidade do infográfico pelo número massivo de dados	Mudanças periódicas no nosso planejamento devido à quantidade de atividades a serem desenvolvidas	Dificuldade na execução de todas as tarefas durante a sprint mais curta.					
	Média	3		Dependência de aulas dos professores para dar continuidade ao projeto	Capacidade insuficiente do computador para processar a quantidade de dados.	Baixo conhecimento técnico individual entre os membros da equipe na arquitetura do projeto					
	Baixa	4				Baixa qualidade dos dados levando a resultados imprecisos e incorretos					
	Muito Baixa	5		Dependência de aulas dos professores para dar continuidade ao projeto	Exposição de dados sensíveis pelo github						
			1	2	3	4	5	5	4	3	2
	Muito Baixa	Baixa	Média	Alta	Muito Alta	Muito Alta	Alta	Média	Baixa	Muito Baixa	
Impacto											

Imagen 8 - Matriz de risco 4

Plano de Ação				
Matriz	Descrição do Impacto		Responsável	Descrição da Ação
3-4	A incapacidade do computador em lidar com a grande quantidade de dados pode resultar em atrasos no pré-processamento, impactando a eficiência da análise e o cronograma do projeto.		Thainá	Garantir o pleno funcionamento da máquina virtual compatível com a utilização para que recursos sejam adequados para o volume de dados do projeto.
3-3	A dificuldade em tornar o infográfico comprehensível pode comprometer a utilidade do projeto, dificultando a análise de informações valiosas.		Vitória	Estar sempre em contato com o professor de UX para obter feedback sobre a usabilidade e com o cliente para saber se suas dores foram solucionadas
2-3	Faltas dos integrantes não avisadas antecipadamente podem acarretar no atraso de tarefas e sobrecarga de outros integrantes		Larissa	Comunicação clara e notificação antecipada em caso de ausência. Adaptação do horário do membro do time que terá faltas recorrentes para que suas atividades sejam cumpridas dentro do prazo da atividade
3-5	Dificuldades técnicas dos integrantes sem aviso prévio pode atrasar o andamento e entrega do projeto além de prejudicar o engajamento do grupo		Gabriela	Garantir que os POs do time estejam a par do problema, a fim de que o professor de programação seja recorrido para auxiliar o time. Além disso, os autoestudos devem ser revisados em primeira instância.
1-5	Não possuir uma conta ativa na AWS impede o desenvolvimento, testes e a implantação de recursos na nuvem, o que compromete a eficiência operacional e a entrega do projeto nos prazos estabelecidos.		Larissa	Solicitar a abertura da conta ao professor, destacando a necessidade crítica dentro do projeto.
1-3	Integração dos dados pode demandar muito tempo e resultar em atrasos significativos na entrega do projeto		Thainá	Ao perceber que o problema está acontecendo, é substancial o estudo de outras ferramentas que realizem o processamento e integração dos dados de forma mais facilitada.
2-5	Dificuldade na execução de todas as tarefas durante a sprint mais curta pode resultar em atrasos no projeto, comprometendo prazos e entregas.		Vitória	Priorização das atividades conforme as entregas e garantir que as mesmas sejam finalizadas na data estabelecida.
4-5	A baixa qualidade dos dados podem afetar a credibilidade do projeto e as conclusões resultadas dos dados.		Ueliton	Aplicação de conceitos matemáticos, como o PCA, que permite a ampliação da qualidade dos dados utilizados e garantir melhores resultados.
5-3	Dependência de aulas dos professores para dar continuidade ao projeto pode resultar em atrasos na progressão do projeto.		Sophia	Conduzir uma avaliação das necessidades individuais de aprendizado de cada membro da equipe e identificar se há necessidade de auto-estudos extras.
5-4	Quebra de licenças e exposição de dados sensíveis pelo github pode resultar em perda de dados e danos à reputação do projeto.		Ueliton	Retirada dos dados da internet com o máximo de urgência e anonimação imediata.
2-4	Mudanças periódicas no nosso planejamento devido à quantidade de atividades pode causar estresse adicional à equipe e dificultar o cumprimento dos prazos.		Vitória	Validação de cada artefato da sprint com o professor responsável através do grupo no slack.
1-4	A exclusão da conta da (AWS) pode resultar na perda irreversível de todos os dados, trabalhos, bases e recursos associados à plataforma.		Thainá	Utilizar os labs de cada integrante e manter uma conta adicional na AWS como backup.

Oportunidades

Até o momento, não foram identificadas oportunidades.

Imagen 9 - Descrição 4

Sprint 5:

Matriz de Risco											
Probabilidade:	Riscos						Oportunidades				
	Muito Alta	1	Integração dos dados demandar muito tempo e atrair atraso de projeto	Perda ou Exclusão da Conta da AWS e seus Dados Associados	Não possuir uma conta ativa na AWS						
	Alta	2	Ausência dos integrantes não dev acarretarem atrasos nas entregas dos integrantes	Mudanças periódicas no planejamento devido à quantidade de atividades a serem desenvolvidas	Encerramento Antecipado da Sprint						
	Média	3	Pouca usabilidade do infográfico pelo numero massivo de cidades	Capacidade insuficiente do computador para processar a quantidade de dados.	Baixo conhecimento técnico individual entre os membros da equipe na arquitetura do projeto						
	Baixa	4			Baixa qualidade dos dados levando a resultados imprecisos e incorretos						
	Muito Baixa	5	Dependência de aulas dos professores para dar continuidade ao projeto	Exposição de dados sensíveis pelo github							
		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											

Imagen 10 - Matriz de risco 5

Plano de Ação			
Matriz	Descrição do Impacto	Responsável	Descrição da Ação
3-4	A incapacidade do computador em lidar com a grande quantidade de dados pode resultar em atrasos no pré-processamento, impactando a eficiência da análise e o cronograma do projeto.	Thainá	Garantir o pleno funcionamento da máquina virtual compatível com a utilização para que recursos sejam adequados para o volume de dados do projeto.
3-3	A dificuldade em tornar o infográfico comprehensível pode comprometer a utilidade do projeto, dificultando a análise de informações valiosas.	Vitória	Estar sempre em contato com o professor de UX para obter feedback sobre a usabilidade e com o cliente para saber se suas dores foram solucionadas
2-3	Faltas dos integrantes não avisadas antecipadamente podem acarretar no atraso de tarefas e sobrecarga de outros integrantes	Larissa	Comunicação clara e notificação antecipada em caso de ausência. Adaptação do horário do membro do time que terá faltas recorrentes para que suas atividades sejam cumpridas dentro do prazo da atividade.
3-5	Dificuldades técnicas dos integrantes sem aviso prévio pode atrasar o andamento e entrega do projeto além de prejudicar o engajamento do grupo	Gabriela	Garantir que os Pôs do time estejam a par do problema, a fim de que o professor de programação seja recorrido para auxiliar o time. Além disso, os autoestudos devem ser revisados em primeira instância.
1-5	Não possuir uma conta ativa na AWS impede o desenvolvimento, testes e a implantação de recursos na nuvem, o que compromete a eficiência operacional e a entrega do projeto nos prazos estabelecidos.	Larissa	Solicitar a abertura da conta ao professor, destacando a necessidade crítica dentro do projeto.
1-3	Integração dos dados pode demandar muito tempo e resultar em atrasos significativos na entrega do projeto	Thainá	Ao perceber que o problema está acontecendo, é substancial o estudo de outras ferramentas que realizem o processamento e integração dos dados de forma mais facilitada.
2-5	Encerramento Antecipado da Sprint tem um impacto significativo na entrega final do projeto pela equipe ter um dia a menos para realizar atividades essenciais como revisar e aprimorar o produto.	Vitória	Priorização das atividades conforme as entregas e garantir que as mesmas sejam finalizadas na data estabelecida.
4-5	A baixa qualidade dos dados podem afetar a credibilidade do projeto e as conclusões resultadas dos dados.	Ueliton	Aplicação de conceitos matemáticos, como o PCA, que permite a ampliação da qualidade dos dados utilizados e garantir melhores resultados.
5-3	Dependência de aulas dos professores para dar continuidade ao projeto pode resultar em atrasos na progressão do projeto.	Sophia	Conduzir uma avaliação das necessidades individuais de aprendizado de cada membro da equipe e identificar se há necessidade de auto-estudos extras.
5-4	Quebra de licenças e exposição de dados sensíveis pelo github pode resultar em perda de dados e danos à reputação do projeto.	Ueliton	Retirada dos dados da internet com o máximo de urgência e anonimação imediata.
2-4	Mudanças periódicas no nosso planejamento devido à quantidade de atividades pode causar estresse adicional à equipe e dificultar o cumprimento dos prazos.	Vitória	Validação de cada artefato da sprint com o professor responsável através do grupo no slack.
1-4	A exclusão da conta da (AWS) pode resultar na perda irreversível de todos os dados, trabalhos, bases e recursos associados à plataforma.	Thainá	Utilizar os labs de cada integrante e manter uma conta adicional na AWS como backup.

Oportunidades

Até o momento, não foram identificadas oportunidades.

Imagen 11 - Descrição 5

3.3 TAM SAM SOM

A análise de tamanho do mercado com foco na receita é importante para o planejamento estratégico de uma empresa. Ela fornece informações essenciais para entender o potencial de um negócio em seu setor específico. Neste caso exploramos os conceitos e números relacionados ao mercado de Business Intelligence no contexto do segmento de varejo alimentar e de serviço alimentar no Brasil, com um enfoque na cidade de São Paulo.

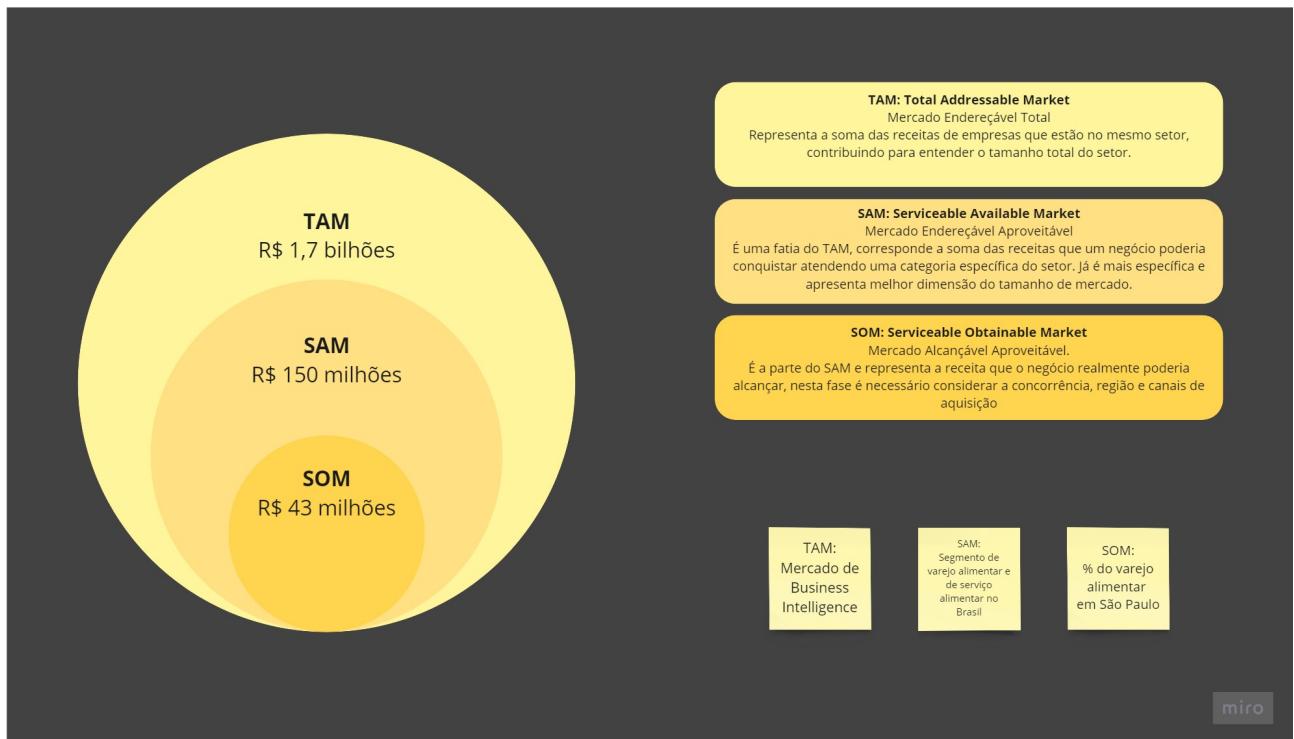


Imagen 12 - TAM/SAM/SOM

TAM - Total Addressable Market (Mercado Endereçável Total) O TAM representa o tamanho total do mercado em termos de receita. No caso do mercado de Business Intelligence no Brasil, o TAM é estimado em R\$1,7 bilhões. Isso inclui todas as receitas geradas por empresas que atuam nesse setor em todo o país. (McKinsey, 2023).

SAM - Serviceable Available Market (Mercado Endereçável Aproveitável) O SAM é uma fatia específica do TAM. Ele corresponde à soma das receitas que um negócio poderia conquistar atendendo a uma categoria específica do setor. No nosso caso, o SAM está relacionado ao segmento de varejo alimentar e de serviço alimentar no Brasil. Estima-se que o SAM seja de R\$150 milhões. Isso leva em consideração que cerca de 40% dessas empresas utilizam soluções de Business Intelligence. (Vivo Meu Negócio, 2023).

SOM - Serviceable Obtainable Market (Mercado Alcançável Aproveitável) O SOM representa a receita que um negócio realmente poderia alcançar. Ele é uma parte do SAM e leva em consideração fatores como a concorrência, a região geográfica e os canais de aquisição. No contexto da indústria alimentícia em São Paulo, o SOM é estimado em R\$43 milhões. Esta é a receita que as empresas que oferecem soluções de Business Intelligence podem efetivamente obter no mercado de varejo alimentar em São Paulo.(Statista, 2023).

4. Análise de Experiência do Usuário

No âmbito empresarial, o entendimento do usuário é um componente fundamental para o sucesso de qualquer empreendimento. Compreender as necessidades, preferências e comportamentos dos usuários não apenas orienta o desenvolvimento de produtos e serviços, mas também influencia diretamente as estratégias de marketing, vendas e a experiência geral do cliente. Nesta seção, exploraremos a importância desse entendimento aprofundado, destacando como ele impacta positivamente as operações e o relacionamento entre a empresa e seus consumidores.

4.1 Personas

Personas é uma técnica comum de pesquisa de mercado e design de experiência do usuário (UX), que tem como função o entendimento das necessidades, desejos, objetivos e comportamentos de um determinado grupo de pessoas. Com esse conhecimento, os desenvolvedores conseguem criar soluções mais adequadas, já que é possível entender melhor as dores, comportamentos e motivações dos usuários.

4.1.1. Persona 1: Juliana Santos

Abaixo tem-se a primeira persona: Juliana Santos, uma profissional de marketing e vendas determinada na empresa de consultoria Integration. Ela enfrenta grande desafios em lidar com grande volume de dados no dia-a-dia da empresa, que por conta da alta quantidade de dados, acaba não conseguindo ter uma boa análise aprofundada, ou seja, há muitas informações que acabam ficando dispersas por conta do alto volume de dados. Sendo assim, há uma grande falta de insights detalhados para suas estratégias. Portanto, suas necessidades incluem acesso a dados detalhados sobre o consumo, insights valiosos para direcionar suas estratégias e acesso rápido a dados relevantes para tomar decisões informadas em tempo real. Logo, deseja ter uma ferramenta que proporcione uma visualização mais personalizada sobre os dados.



JULIANA SANTOS

CONSULTOR DE MARKETING E VENDAS

NOME: Juliana Santos

IDADE: 35 anos

GÊNERO: Feminino

OCUPAÇÃO: Consultora de marketing e vendas

BIOGRAFIA

Juliana Santos é uma profissional de marketing e vendas experiente com mais de uma década de experiência nessa área. Ela trabalha na Integration Consulting, onde se destaca por desenvolver estratégias de mercado inovadoras para os clientes. Juliana é conhecida por sua paixão por dados e análises, que usa para tomar decisões estratégicas informadas, além disso, é orientada por resultados e foco no cliente.

CARACTERÍSTICAS

- Comunicativa;
- Criativa;
- Habilidade analítica;
- Visão estratégica;

DORES

- Dificuldade em lidar com informações dispersas e falta de insights detalhados para direcionar estratégias de marketing e vendas eficazes.
- Falta de acesso rápido a dados relevantes para tomar decisões informadas.

FORMAÇÃO

- SUPERIOR EM MARKETING E PROPAGANDA PELA FACULDADE ESPM;
- MBA EM GESTÃO DE NEGÓCIOS.

NECESSIDADES

- Precisa de acesso a dados detalhados sobre o consumo em diferentes categorias e canais, de forma a compreender o mercado em profundidade;
- Obter insights valiosos obtidos a partir da análise de dados, que podem apoiar suas estratégias de marketing e vendas;
- Precisa de acesso rápido a dados relevantes, permitindo tomar decisões em tempo real para aproveitar oportunidades ou responder a desafios de mercado.

Imagen 13 - Persona 1

Dores:

- Dificuldade em lidar com informações dispersas e falta de insights detalhados para direcionar estratégias de marketing e vendas eficazes.
- Falta de acesso rápido a dados relevantes para tomar decisões informadas.

Necessidades:

- Precisa de acesso a dados detalhados sobre o consumo em diferentes categorias e canais, de forma a compreender o mercado em profundidade;
- Obter insights valiosos obtidos a partir da análise de dados, que podem apoiar suas estratégias de marketing e vendas;
- Precisa de acesso rápido a dados relevantes, permitindo tomar decisões em tempo real para aproveitar oportunidades ou responder a desafios de mercado.

Nível de letramento digital:

Juliana possui um nível médio de letramento digital, demonstra habilidade e familiaridade com tecnologias e ferramentas de visualização de dados digitais. Sua experiência a capacita a interagir de maneira fácil com o nosso sistema e plataforma online que será personalizada para a área em que ela atua na empresa.

Cenários de Interação:

- Reuniões estratégicas:

Juliana interage com o sistema durante reuniões estratégicas, onde precisa acessar dados em tempo real para respaldar suas decisões.

- Análise de desempenho:

Durante a análise de desempenho de campanhas, Juliana utiliza a visualização dos infográficos para obter insights detalhados e ajustar as estratégias de marketing conforme necessário.

- Planejamento de Campanhas:

Ao planejar novas campanhas, Juliana utiliza a ferramenta para analisar dados de consumo em diferentes categorias e canais, fundamentando suas estratégias.

- Resposta a Mudanças de Mercado:

Em situações de mudanças rápidas no mercado, Juliana interage com o sistema para tomar decisões imediatas com base em dados relevantes e atualizados.

Exemplos de uso e preferências:

Juliana adota uma abordagem prática, explorando a plataforma para visualizar dashboards/infográficos intuitivos e personalizados. Ela foca em analisar dados de

consumo, adaptando a interface para decisões estratégicas rápidas e informadas. A agilidade no acesso a dados é crucial para sua orientação por resultados e foco no cliente.

Citações a partir dos pensamentos de Juliana:

- "A capacidade de acesso rápido a dados detalhados é crucial para o sucesso de nossas estratégias."
- "A análise de dados impulsiona cada decisão que tomamos."

4.1.2. Persona 2: Augusto Ribeiro

Abaixo tem-se a segunda persona: Augusto sempre teve uma ótima visão de negócio e se aprimorou ainda mais após ter iniciado sua jornada na Integration Consulting. Porém com o passar do tempo ele notou que sentia falta de estar mais próximo da área de tecnologia. Atualmente está atuando na equipe de Tech & Digital onde ele consegue juntar seus conhecimentos de negócio criando novas ferramentas tecnológicas e ajustar as já existentes que auxiliam o trabalho dos consultores da Integration.



AUGUSTO RIBEIRO

TECH & DIGITAL

NOME: Augusto Ribeiro

IDADE: 32 anos

GÊNERO: Masculino

OCUPAÇÃO: Tech & Digital

BIOGRAFIA

Augusto sempre teve uma ótima visão de negócio e se aprimorou ainda mais após ter iniciado sua jornada na Integration Consulting. Porém com o passar do tempo ele notou que sentia falta de estar mais próximo da área de tecnologia. Atualmente está atuando na equipe de Tech & Digital onde ele consegue juntar seus conhecimentos de negócio criando novas ferramentas tecnológicas e ajustar as já existentes que auxiliam o trabalho dos consultores da Integration.

CARACTERÍSTICAS

- Comunicativo;
- Detalhista;
- Observador;
- Inquieto;

DORES

- Gasto de tempo em atividades repetitivas;
- Desorganização pelo uso de muitas planilhas;
- Falta de tempo para automatizar processos;
- Pouca eficiência e sofisticação ao criar soluções para consultores de Marketing e vendas;
- Abandono de dados na criação de soluções;
- Consultores com pouco contato com tecnologias de dados;

FORMAÇÃO

- SUPERIOR EM MARKETING E GESTÃO DE VENDAS PELA FGV
- PÓS EM ENGENHARIA DA COMPUTAÇÃO PELA MAUÁ;

NECESSIDADES

- Aumentar produtividade ao automatizar processos repetitivos;
- Diminuir o uso de planilhas de Excel;
- Aumentar a robustez ao criar soluções para os consultores;
- Facilitar a procura de dados;
- Criar interface didática para uso dos consultores;
- Flexibilidade da ferramenta para uso com diferentes tipos de dados;
- Compatibilidade com outras ferramentas tecnológicas;

Imagen 14 - Persona 2

Dores:

- Gasto de tempo em atividades repetitivas;
- Desorganização pelo uso de muitas planilhas;
- Falta de tempo para automatizar processos;

- Pouca eficiência e sofisticação ao criar soluções para consultores de Marketing e vendas;
- Abandono de dados na criação de soluções;
- Consultores com pouco contato com tecnologias de dados;

Necessidades:

- Aumentar produtividade ao automatizar processos repetitivos;
- Diminuir o uso de planilhas de Excel;
- Aumentar a robustez ao criar soluções para os consultores;
- Facilitar a procura de dados;
- Criar interface didática para uso dos consultores;
- Flexibilidade da ferramenta para uso com diferentes tipos de dados;
- Compatibilidade com outras ferramentas tecnológicas;

Nível de letramento digital:

Augusto possui um nível avançado de letramento digital, demonstrando habilidades sofisticadas no uso de tecnologias e ferramentas digitais.

Cenários de Interação:

- Desenvolvimento de ferramentas:

Augusto interage com o sistema durante o desenvolvimento de novas ferramentas tecnológicas, buscando aumentar a eficiência e sofisticação para os consultores.

- Automatização de processos:

Ao lidar com atividades repetitivas, Augusto utiliza a plataforma para automatizar processos, economizando tempo e recursos.

- Integração com outras ferramentas:

Augusto interage com o sistema para garantir a compatibilidade e integração eficiente com outras ferramentas tecnológicas utilizadas na Integration.

- Treinamento de consultores:

Com a criação de infográficos e automatização de dados, Augusto facilita o treinamento dos consultores no uso das ferramentas, garantindo uma adoção mais eficaz.

Exemplos de uso e preferências:

Augusto prefere uma abordagem hands-on ao desenvolver soluções, utilizando a plataforma para criar dashboards intuitivos e funcionalidades personalizadas para facilitar o trabalho dos consultores.

Citações a partir dos pensamentos de Augusto:

- "A flexibilidade para lidar com diferentes clientes é crucial. Precisamos de soluções que se adaptem às necessidades específicas de cada projeto."
- "Automatizar processos é fundamental para que eu possa focar em inovação e melhorias constantes."
- "Uma interface intuitiva é o segredo para o sucesso da adoção das ferramentas pelos consultores."

4.2 Jornada do Usuário

A Jornada do Usuário é uma representação visual das etapas que um usuário percorre ao interagir com um produto ou serviço, incluindo desde o primeiro contato com a solução até a conclusão de suas tarefas ou objetivos. Ela descreve o passo a passo percorrido, detalhando todos os pontos de contato e interações do ponto de vista do usuário, seus sentimentos e sensações em cada fase. Por mapear realmente como é a experiência do usuário, ela é essencial para identificar áreas de melhoria e criar soluções mais intuitivas e centradas no usuário.

Para uma melhor visualização acessar o link:

https://miro.com/app/board/uXjVNU1a7B0=?share_link_id=523674542118

4.2.1. Jornada do usuário persona 1 - Juliana Santos, Consultora de Marketing e Vendas.

 Juliana Santos Consultor de Marketing e Vendas • 35 anos - Feminino  Journey goal										
		Avaliação da solução	Utilização Inicial da ferramenta	Tomada de decisão	Monitoramento e utilização no dia-a-dia do trabalho					
Fases da jornada	Conferir a solução do pipeline de Big Data baseada em recursos da AWS para análise de dados									
Ações	Juliana conhece a solução de análise de dados baseada na AWS para aprimorar suas estratégias de marketing e vendas	Pesquisar mais a fundo a solução, explorando seus recursos e funcionalidades	Juliana se inscreve na ferramenta e começa a usar para análise de dados	Juliana precisa fazer uma análise de dados para ter um melhor direcionamento. Logo, realiza a visualização dos infográficos.	Receber atualizações da ferramenta, com dados novos e assim ganhar mais conhecimento a partir dos dados					
Pensamentos	"Esta solução pode ser a resposta para as minhas necessidades de análise de dados e obtenção de insights."	"Vou garantir que essa solução atenda às minhas necessidades específicas."	"Essa ferramenta realmente me trará grandes benefícios para aprimorar ainda mais as minhas decisões."	"Nossa, como essa visualização está facilitando no meu dia-a-dia e aprimorando as minhas estratégias."	"Esta ferramenta fornece ótimas visualizações que tem agregado muito. Não preciso mais perder tanto tempo fazendo manualmente a análise de dados."					
Sentimentos	Alta expectativa	Interesse	Responabilidade	Curiosidade	Empolgação	Confiança	Satisfação	Orgulho	Realização	Felicidade
Oportunidades	Utilizar uma solução que processa um grande volume de dados para geração de insights	Ter uma ferramenta que auxilie na tomada de decisão no dia-a-dia	Ter novas descobertas com essa ferramenta e assim, obter insights e tomar decisões mais embasadas	Aprimoramento dentro das estratégias de marketing e vendas. Além de maior entendimento e maior compreensão sobre os dados da empresa	Dar os dados organizados, e assim, com esse valor, poderá conseguir resultados melhores, mais eficazes, e, logo, resultados melhores e mais relevantes					
Touchpoints	AWS	AWS	Acesso aos infográficos e cubo de dados	Infográfico Digital	Infográfico Digital					
Final da jornada	A partir do desenvolvimento do cubo de dados e infográficos para uma fácil visualização dos dados, a Juliana, no cargo de marketing dentro da consultoria, conseguirá ter uma ferramenta que a permita saber o potencial de consumo de cada categoria, região e canais. Sendo assim, com a utilização dessa ferramenta no dia a dia, surgirão mais insights para um melhor direcionamento de suas estratégias e decisões dentro da empresa, buscando sempre aumentar seus resultados e ser mais assertiva tendo uma maior compreensão sobre os clientes da Integration.									
Feedbacks e insights	Economia de tempo	Automatização de processos	Eficiência nas atividades	Insights relevantes a partir da visualização dos dados	Aprimorar estratégias de Marketing					

Imagen 15 - Jornada 1

4.2.2. Jornada do usuário persona 2 - Augusto Ribeiro, Tech Digital

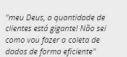
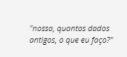
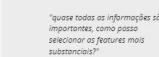
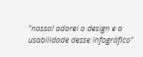
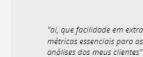
 Augusto Ribeiro Consultor Tech & Digital • 47 anos - Masculino					
 Que o consultor tenha as informações necessárias para ser visualizadas por meio de um infográfico					
 Fases da jornada	 Coletar os dados de potenciais clientes	 Coletar de dados do governo	 Integrar os dois arquivos contendo a persistência no projeto	 Visualizar um infográfico com as informações necessárias	 Tomar decisões a partir dos achados e fazer a troca de conhecimento
 Ações	 Verificar a partir da API do cubo as informações corretas e que são do cliente.	 Colocar os arquivos PDF e de CAPI em um formato correto mais amigável script Python	 Colocar os arquivos diálogos da API e de CAPI para que sejam integrados ao banco de dados.	 Organizar os arquivos, da maneira correta, da maneira que é mais correspondente a todos os processos e que é mais fácil de se integrar ao banco de dados.	 Os sempre foi desafiador, pois existem, na internet, muitos resultados que correspondem a APIs diferentes, que não conseguem integrar os dados para a troca de dados integrados.
 Pensamentos	 "meu Deus, a quantidade de clientes está gigante! Não sei como vou fazer o coleta de dados de forma eficiente"	 "nossa, quantos dados antigos, o que eu faço?"	 "quase todas as informações são importantes, como posso selecionar os features mais substanciais?"	 "nossa! adorei o design e a usabilidade desse infográfico"	 "oi, que facilidade em extrair métricas essenciais para as análises dos meus clientes!"
 Sentimentos	 Feliz por ter realizado a prospecção de muitos clientes	 Desesperado pela quantidade de dados	 Satisfeito por poder incorporar dados reais em sua análise	 Confuso pela necessidade de escolher quais são mais importantes para a empresa	 Animado por poder selecionar os dados mais importantes para a empresa
 Oportunidades	 Otimizar a eficiência e reduzindo o tempo de execução da API para melhorar a eficiência.	 Formar tutoriais de uso para ajudar os usuários a usar o script	 Monitorar e realizar reviews sobre mudanças na API externa	 Melhorar as ferramentas que auxiliam a identificação do problema e otimizar as ferramentas mais produtivas dos usuários.	 Permitir que os usuários personalizem a integração de diferentes tipos de necessidades específicas.
 Touchpoints					
 Final da jornada	A partir da criação do pipeline e facilitação da visualização dos dados, o Augusto consegue otimizar o processo de captação, associação e tomada de decisão baseado em dados, para outros membros da equipe.				
 Feedback e Insights	 Automatização de processos	 Auxiliar o profissional a entender os processos governamentais, econômicos e sociais para prever as tendências futuras.	 Gerenciamento de dados contínuos	 Identificar as as tendências presentes no projeto e selecionar para utilização prática para análise.	 Recomendar o tipo de infográfico mais adequado para a necessidade de cada usuário e fornecer opções de personalização.

Imagen 16 - Jornada 2

4.3 User Stories

User stories desempenham um papel crucial na construção de projetos. Elas oferecem uma abordagem centrada no usuário para definir funcionalidades, garantindo que as soluções desenvolvidas sejam verdadeiramente relevantes e atendam às necessidades reais dos usuários. Ao utilizar linguagem simples e direta, as user stories facilitam a comunicação entre todos os envolvidos no projeto, desde desenvolvedores até stakeholders, criando um entendimento compartilhado do que precisa ser alcançado. Esta clareza ajuda a evitar mal-entendidos e retrabalhos, otimizando o processo de desenvolvimento. Além disso, por representarem as funcionalidades do sistema, elas guiam a organização da equipe em relação às Sprints, sendo divididas em tarefas no quadro Kan Ban e produzidas dentro de uma sprint. Dessa maneira, foram desenvolvidas 12 user stories para as personas descritas acima: Consultor de Marketing e Vendas e Consultor de Tech&Digital.

Número	User story 1
Persona	Consultor de marketing e vendas
História	Eu, como consultor de marketing e vendas, quero visualizar dados sobre a localização geográfica dos nossos potenciais clientes, seus segmentos de mercado e informações sobre concorrentes.
Critérios de Aceitação	<ol style="list-style-type: none">1. O sistema deve categorizar clientes por localização geográfica (cidade, estado, país) e segmento de mercado.2. O sistema deve fornecer gráficos dos canais de venda com base nos dados de clientes.
Testes de Aceitação	<ol style="list-style-type: none">1. Verificar se o sistema destaca os canais de venda ao inserir dados de clientes.

	2. Confirmar a categorização automática de um novo cliente no painel de controle.
--	---

Descrição User Story 1

- **Número:** 1
- **Persona:** Consultor de marketing e vendas
- **História:** Visualizar dados geográficos, segmentação de clientes e informações sobre concorrentes.
- **Small (Pequena):** é pequena para ser realizada em uma iteração, pois se concentra na visualização de dados específicos para um consultor de marketing e vendas.
- **Independent (Independente):** é relativamente independente, pois se concentra em requisitos específicos para categorização e visualização de dados de clientes.
- **Negotiable (Negociável):** é negociável, pois oferece flexibilidade para ajustar os requisitos específicos relacionados à categorização e visualização de dados.
- **Testabilidade:** é clara com testes de inserção de dados e verificação de categorização automática.
- **Priorização:** a importância é média.
- **Relações e Dependências:** a User Story 1 está relacionada à User Story 9, pois ambas envolvem a categorização de clientes com base em comportamentos de compra.

Número	User story 2
Persona	Consultor de Tech&Digital
História	Eu, como consultor de Tech&Digital, quero integrar uma solução automatizada para acessar e analisar dados rapidamente.
Critérios de Aceitação	<p>1. O sistema deve oferecer uma interface intuitiva que permita a visualização de dados através de dashboards interativos e outros formatos relevantes.</p> <p>2. O sistema deve mostrar métricas de desempenho do processo de coleta de dados.</p> <p>3. O sistema deve demonstrar uma redução significativa no tempo necessário para buscar e processar os dados, quando comparado ao método anterior.</p>
Testes de Aceitação	<p>1. Comparar o tempo de coleta e análise de dados antes e depois da implementação.</p> <p>2. Confirmar a visualização de dados através de dashboards.</p> <p>3. Verificar as métricas de desempenho apresentadas pelo sistema.</p>

Descrição User Story 2

Número: 2

Persona: Consultor de Tech&Digital

História: Integrar solução automatizada para acessar e analisar dados rapidamente.

- **Small (Pequena):** a história parece abrangente, mas os critérios de aceitação podem ser divididos em tarefas menores, realizando assim em uma iteração.
- **Independent (Independente):** pode ser independente se as interfaces e métricas forem desenvolvidas de maneira modular, permitindo a implementação gradual.
- **Negotiable (Negociável):** é negociável, pois os requisitos podem ser ajustados com base nas necessidades específicas de visualização e desempenho.
- **Testabilidade:** testes de interface intuitiva, métricas de desempenho e redução de tempo.
- **Priorização:** a importância é média.
- **Relações e Dependências:** a User Story 2 pode depender da implementação bem-sucedida da User Story 10, que envolve algoritmos avançados e automação.

Número	User story 3
Persona	Consultor Tech&Digital
História	Eu, como consultor Tech&Digital, quero um sistema que me forneça uma compreensão clara e objetiva dos propósitos e metas do projeto, para que possa alinhar minhas análises e decisões de forma mais eficaz.
Critérios de Aceitação	<p>1. O sistema deve integrar uma funcionalidade de automatização na coleta e leitura de dados, evitando retrabalhos e falhas manuais.</p> <p>2. O sistema deve ser adaptável e aceitar diferentes formatos de dados do cliente, permitindo uma inserção sem complicações.</p> <p>3. A plataforma deve possuir um módulo ou seção dedicada a apresentar, de forma clara e concisa, os objetivos e metas do projeto.</p>
Testes de Aceitação	<p>1. Ao acessar o sistema, o consultor deve ser capaz de identificar facilmente os objetivos e metas do projeto sem necessidade de busca extensiva.</p> <p>2. Ao inserir diversos formatos de dados do cliente, o sistema deve processá-los sem erros e de forma eficiente.</p> <p>3. Ao ativar a funcionalidade de automatização, o sistema deve realizar a coleta e leitura de dados sem intervenção manual e apresentar os resultados em tempo hábil.</p>

Descrição User Story 3

Número: 3

Persona: Consultor Tech&Digital

História: Sistema claro e objetivo para compreensão de propósitos e metas do projeto.

- **Small (Pequena):** a história é relativamente grande, mas pode ser dividida em tarefas menores, como implementar a automatização da coleta de dados e a adaptação para diferentes formatos.
- **Independent (Independente):** pode ser independente se as funcionalidades de automatização e adaptação de dados forem desenvolvidas de forma separada.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à automatização e adaptação de dados.
- **Testabilidade:** testes de automatização, adaptação de dados e apresentação clara de objetivos e metas do projeto.
- **Priorização:** a importância pode ser média a alta.
- **Relações e Dependências:** a User Story 3 está relacionada à User Story 10, pois ambas envolvem automação e atualização de dados.

Número	User story 4
Persona	Consultor Tech&Digital
História	Como consultor Tech&Digital, desejo um sistema que priorize a segurança e a conformidade dos dados, garantindo que as informações sensíveis do cliente não sejam armazenadas no cubo de dados, a fim de manter a integridade e cumprir com as regulamentações de governança de dados.
Critérios de Aceitação	1. Qualquer dado sensível ou identificável do cliente inserido no sistema não deve ser retido ou armazenado no cubo de dados sem criptografia.
Testes de Aceitação	1. Ao inserir dados sensíveis do cliente, realizar uma busca no cubo de dados após um período estabelecido e confirmar a criptografia desses dados.

Descrição User Story 4

Número: 4

Persona: Consultor Tech&Digital

História: Priorizar segurança e conformidade dos dados, evitando armazenamento no cubo de dados.

- **Small (Pequena):** é pequena, pois se concentra principalmente na segurança e conformidade dos dados.
- **Independent (Independente):** é independente, pois trata de requisitos de segurança sem depender de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos de segurança e conformidade.
- **Testabilidade:** testes de inserção de dados sensíveis e busca criptografada.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 4 está relacionada à User Story 8, pois ambas envolvem autenticação e segurança dos dados.

Número	User story 5
Persona	Consultor de Tech&Digital
História	Eu, como consultor de Tech&Digital, quero que o sistema acelere a busca e análise de dados no processo "go to market".
Critérios de Aceitação	<p>1. O sistema deve possuir algoritmos avançados para análises rápidas sem comprometer a precisão.</p> <p>2. A plataforma deve fornecer resultados de análise em no máximo 3 minutos.</p> <p>3. A solução deve incluir ferramentas de automação que reduzam o esforço manual na coleta e processamento de dados.</p>
Testes de Aceitação	<p>1. Comparar o tempo médio gasto na busca e análise de dados antes e após a implementação do sistema.</p> <p>2. Submeter um conjunto de dados ao sistema e verificar o tempo necessário para receber os resultados.</p> <p>3. Utilizar a automação do sistema para coletar e processar dados e monitorar a eficiência em relação ao método manual anterior.</p>

Descrição User Story 5

Número: 5

Persona: Consultor de Tech&Digital

História: Acelerar busca e análise de dados no processo "go to market".

- **Small (Pequena):** é razoavelmente grande, mas pode ser dividida em tarefas menores, como a implementação de algoritmos avançados, ferramentas de automação e limites de tempo.
- **Independent (Independente):** pode ser independente se cada componente (algoritmos, automação, limites de tempo) for implementado de forma separada.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à análise rápida, automação e tempo de resposta.
- **Testabilidade:** a testabilidade é clara com testes de algoritmos, tempo de resposta e ferramentas de automação.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 5 está relacionada à User Story 9, pois ambas envolvem tempo de resposta e eficiência na análise de dados.

Número	User story 6
Persona	Consultor de Tech&Digital
História	Como consultor Tech&Digital, quero que o sistema suporte a integração nativa com plataformas de visualização como PowerBI e permita a execução e aplicação de scripts Python, possibilitando análises mais aprofundadas e visualizações personalizadas dos dados.
Critérios de Aceitação	<p>1. O sistema deve possuir uma API ou interface que permita a fácil integração com PowerBI e outras ferramentas de visualização.</p> <p>2. A plataforma deve ter um ambiente seguro onde scripts Python possam ser carregados e executados.</p> <p>3. As visualizações e análises criadas através dessas integrações devem ser precisas e refletir os dados reais do sistema.</p>
Testes de Aceitação	<p>1. Importar um conjunto de dados para o PowerBI a partir do sistema e criar uma visualização para confirmar a integração.</p> <p>2. Carregar e executar um script Python que processe e analise um conjunto de dados do sistema.</p> <p>3. Comparar os resultados da análise e visualização criados através das integrações com os dados originais do sistema para garantir precisão.</p>

Descrição User story 6

Número: 6

Persona: Consultor Tech&Digital

História: Suporte à integração nativa com PowerBI e execução de scripts Python.

- **Small (Pequena):** é relativamente grande, mas pode ser dividida em tarefas menores, como:
 - a implementação da API;
 - ambiente seguro para scripts Python e;
 - integração com ferramentas de visualização.
- **Independent (Independente):** pode ser independente se cada componente (API, ambiente seguro, integração) for implementado de forma separada.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à integração com ferramentas externas e execução de scripts Python.
- **Testabilidade:** testes de integração com PowerBI, execução de scripts Python e precisão das visualizações.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 8 está relacionada à User Story 11, pois ambas envolvem integração e interoperabilidade com diferentes plataformas.

Número	User story 7
Persona	Consultor de Marketing e Vendas
História	Como consultor de marketing e vendas, quero que o sistema identifique e categorize onde os potenciais clientes mais frequentemente realizam suas compras, permitindo uma compreensão detalhada dos canais de venda, áreas geográficas e segmentos de cliente para que possamos ajustar e focar nossa estratégia de marketing de forma mais eficaz.
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve coletar e processar dados de compras de potenciais clientes. 2. Deve oferecer visualizações geográficas claras que destaquem áreas de alta concentração de vendas. 3. O sistema deve ser capaz de segmentar clientes com base em comportamentos de compra e categorias de produtos. 4. Precisa identificar, listar e classificar os canais de venda mais populares entre os potenciais clientes.
Testes de Aceitação	<ol style="list-style-type: none"> 1. Inserir dados de compras e verificar se o sistema categoriza corretamente os canais de venda. 2. Visualizar um mapa que destaca as áreas geográficas com maior concentração de compras. 3. Analisar segmentos de clientes para confirmar a correta segmentação baseada em comportamento e categorias de produtos.

Descrição User story 7

Número: 7

Persona: Consultor de marketing e vendas

História: Identificar e categorizar locais de compra de potenciais clientes para ajustar estratégias de marketing.

- **Small (Pequena):** é relativamente grande, envolvendo coleta, processamento e visualização de dados de compras. Pode ser dividida em tarefas menores.
- **Independent (Independente):** pode ser independente se cada componente (coleta, processamento, visualização) for implementado de forma separada.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à categorização e segmentação de clientes.
- **Testabilidade:** testes de categorização e segmentação de clientes com base em comportamentos de compra.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 7 está relacionada à User Story 1, pois ambas envolvem a categorização de clientes.

Número	User story 8
Persona	Consultor de Tech&Digital
História	Como consultor de Tech&Digital, quero garantir a autenticação adequada no sistema para proteger os dados sensíveis dos clientes.
Critérios de Aceitação	<p>1. O sistema deve exigir autenticação para acesso.</p> <p>2. Os dados do cliente não devem ser armazenados permanentemente no storage após a análise.</p>
Testes de Aceitação	<p>1. Tentar acessar o sistema sem autenticação e verificar se o acesso é negado.</p> <p>2. Verificar se os dados do cliente são removidos automaticamente do armazenamento após a análise ser concluída.</p>

Descrição User Story 8

Número: 8

Persona: Consultor de Tech&Digital

História: Garantir autenticação adequada no sistema para proteger dados sensíveis dos clientes.

- **Small (Pequena):** é pequena, pois se concentra principalmente na autenticação e também na não permanência dos dados no armazenamento.
- **Independent (Independente):** é independente, pois trata de requisitos específicos de autenticação sem depender de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à segurança e armazenamento de dados.
- **Testabilidade:** testes de autenticação e remoção automática de dados sensíveis.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 8 está relacionada à User Story 4, pois ambas envolvem autenticação e segurança dos dados.

Número	User story 9
Persona	Consultor de Tech&Digital
História	Como consultor de Tech&Digital, quero garantir que o sistema seja ágil e eficiente na coleta e análise de dados, permitindo-nos tomar decisões em tempo hábil.
Critérios de Aceitação	<p>1. O sistema deve processar solicitações em menos de 5 minutos.</p> <p>2. As análises de dados devem ser concluídas em um tempo definido.</p>
Testes de Aceitação	<p>1. Medir o tempo de resposta do sistema durante diferentes solicitações e analisar se passam de 5 minutos.</p> <p>2. Medir o tempo necessário para a conclusão das análises de dados e verificar se passa de 5 minutos.</p> <p>3. Avaliar a precisão dos dados coletados e processados.</p>

Descrição User Story 9

Número: 9

Persona: Consultor de Tech&Digital

História: Sistema ágil e eficiente na coleta e análise de dados para tomada de decisões.

- **Small (Pequena):** é pequena, pois se concentra principalmente no tempo de resposta e conclusão das análises de dados.
- **Independent (Independente):** é independente, pois trata de requisitos específicos de desempenho sem depender de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados ao tempo de resposta e conclusão das análises.
- **Testabilidade:** testes de tempo de resposta, conclusão de análises e precisão dos dados.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 9 está relacionada à User Story 5, pois ambas envolvem tempo de resposta e eficiência na análise de dados.

Número	User story 10
Persona	Consultor Tech&Digital
História	Como consultor de Tech&Digital, desejo ter a capacidade de atualizar automaticamente os dados sempre que o governo lançar novos dados, assegurando que nossas análises estejam sempre atualizadas.
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve permitir a inserção automática de novos dados. 2. O sistema deve validar a integridade dos dados inseridos.
Testes de Aceitação	<ol style="list-style-type: none"> 1. Inserir novos dados automaticamente e verificar se são aceitos. 2. Tentar inserir dados inválidos e garantir que o sistema os recuse.

Descrição User Story 10

Número: 10

Persona: Consultor de Tech&Digital

História: Atualização automática de dados com lançamentos governamentais para manter análises sempre atualizadas.

- **Small (Pequena):** é pequena, pois se concentra na inserção automática e validação de novos dados.
- **Independent (Independente):** é independente, pois trata de requisitos específicos de atualização de dados sem depender de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, pois permite ajustes nos requisitos relacionados à inserção automática e validação de dados.
- **Testabilidade:** testes de inserção automática e validação de novos dados.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 12 está relacionada à User Story 3, pois ambas envolvem automação e atualização de dados.

Número	User story 11
Épico	Integração e Interoperabilidade
Persona	Consultor de Tech&Digital
História	Como consultor de Tech&Digital, desejo que o sistema seja compatível com diferentes plataformas e bancos de dados, a fim de maximizar a flexibilidade e eficiência na coleta e análise de dados.
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve ser compatível com diferentes bancos de dados NoSQL e SQL. 2. O sistema deve garantir a integridade dos dados ao integrar com diferentes plataformas..
Testes de Aceitação	<ol style="list-style-type: none"> 1. Integrar o sistema com diferentes bancos de dados e verificar a compatibilidade. 2. Verificar a integridade dos dados após a integração com diferentes plataformas.

Descrição User Story 11

Número: 11

Persona: Consultor de Tech&Digital

História: Sistema compatível com diferentes plataformas e bancos de dados.

- **Small (Pequena):** é de tamanho moderado, envolvendo a compatibilidade com bancos de dados NoSQL e SQL, bem como a integridade dos dados ao integrar com diferentes plataformas.
- **Independent (Independente):** é independente, pois trata da compatibilidade e integridade dos dados, sem depender diretamente de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, permitindo ajustes nos requisitos relacionados à compatibilidade com diferentes bancos de dados e plataformas.
- **Testabilidade:** testar o banco de dados, a integridade dos dados e também a integração das plataformas.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 11 está relacionada à User Story 3, pois ambas envolvem automação e atualização de dados.

Número	User story 12
Persona	Consultor de Tech&Digital
História	Como consultor de Tech&Digital, quero que o sistema permita a migração de um serviço cloud para outro de forma simples e eficiente, para garantir flexibilidade e otimização de custos conforme as necessidades da empresa.
Critérios de Aceitação	<ol style="list-style-type: none"> 1. O sistema deve suportar a integração com múltiplos provedores de serviços cloud. 2. O sistema deve garantir a integridade e segurança dos dados durante a migração. 3. A migração deve ser realizada sem interrupções significativas ou tempo de inatividade.
Testes de Aceitação	<ol style="list-style-type: none"> 1. Integrar o sistema com diferentes provedores de serviços cloud e confirmar a compatibilidade. 2. Confirmar a integridade dos dados após a migração. 3. Monitorar o tempo de inatividade durante a migração e garantir que esteja dentro de um limite aceitável

Descrição User Story 12

Número: 12

Persona: Consultor de Tech&Digital

História: Sistema que permite a migração eficiente entre serviços cloud.

- **Small (Pequena):** é de tamanho moderado, abrangendo a integração com múltiplos provedores de serviços cloud, garantia de integridade e segurança dos dados durante a migração, e a realização da migração sem interrupções significativas ou tempo de inatividade.
- **Independent (Independente):** é independente, pois trata da migração entre serviços cloud sem depender diretamente de outras funcionalidades.
- **Negotiable (Negociável):** é negociável, permitindo ajustes nos requisitos relacionados à integração com provedores de serviços cloud, garantia de integridade e tempo de inatividade durante a migração.
- **Testabilidade:** testar a migração dos dados com serviços clouds.
- **Priorização:** alta importância.
- **Relações e Dependências:** a User Story 11 está relacionada às seguintes User Storys: 10 e 11, pois ambas envolvem automação, atualização de dados e testes com os dados.

4.4.Wireframe

O processo de desenvolvimento de um dashboard eficaz é essencial para proporcionar uma experiência de usuário intuitiva e informativa. Antes de qualquer codificação ou design final, a etapa inicial envolve a criação de wireframes. Estes esboços esquemáticos servem como uma representação visual preliminar do layout, estrutura e disposição de gráficos e infográficos no dashboard. Nesta seção, apresentaremos o nosso wireframe antes do nosso design final, o qual possui um design que mostra quais serão os principais dados para o desenvolvimento do nosso projeto.

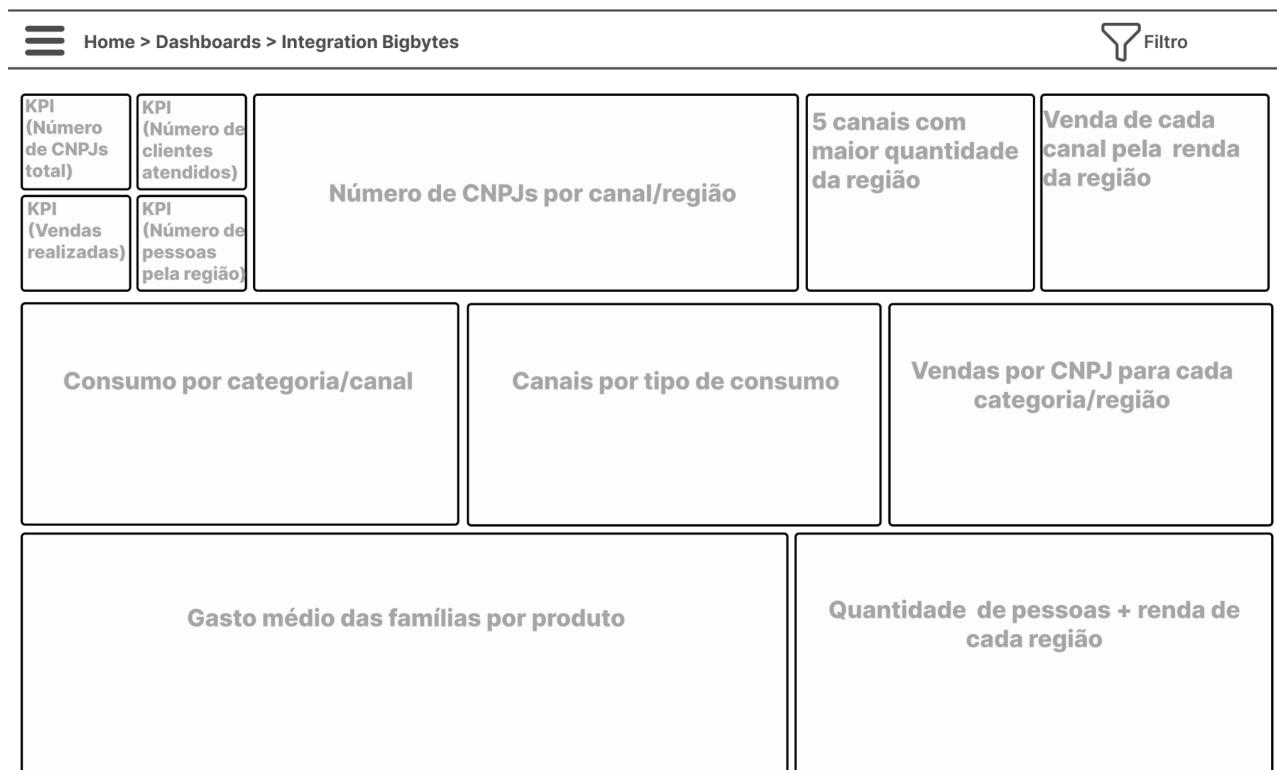


Imagen 17 - wireframe 1

Estado	Cidade	Bairro	1990-2023
Todos os estados Acre (AC) Alagoas (AL) Amapá (AP) Amazonas (AM) Bahia (BA) Ceará (CE) Distrito Federal (DF) Espírito Santo (ES) Goiás (GO) Maranhão (MA)			1940 2023
Selecionar canais:		Selecionar categorias:	
<input type="checkbox"/> mercado	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> restaurante	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Selecionar categoria(CNAE):		Selecionar tipos de consumo:	
<input type="checkbox"/>		<input type="checkbox"/>	
<input type="checkbox"/>		<input type="checkbox"/>	
<input type="checkbox"/>		<input type="checkbox"/>	
<input type="checkbox"/>		<input type="checkbox"/>	
Limpar		Confirmar	

Imagen 18 - wireframe 2

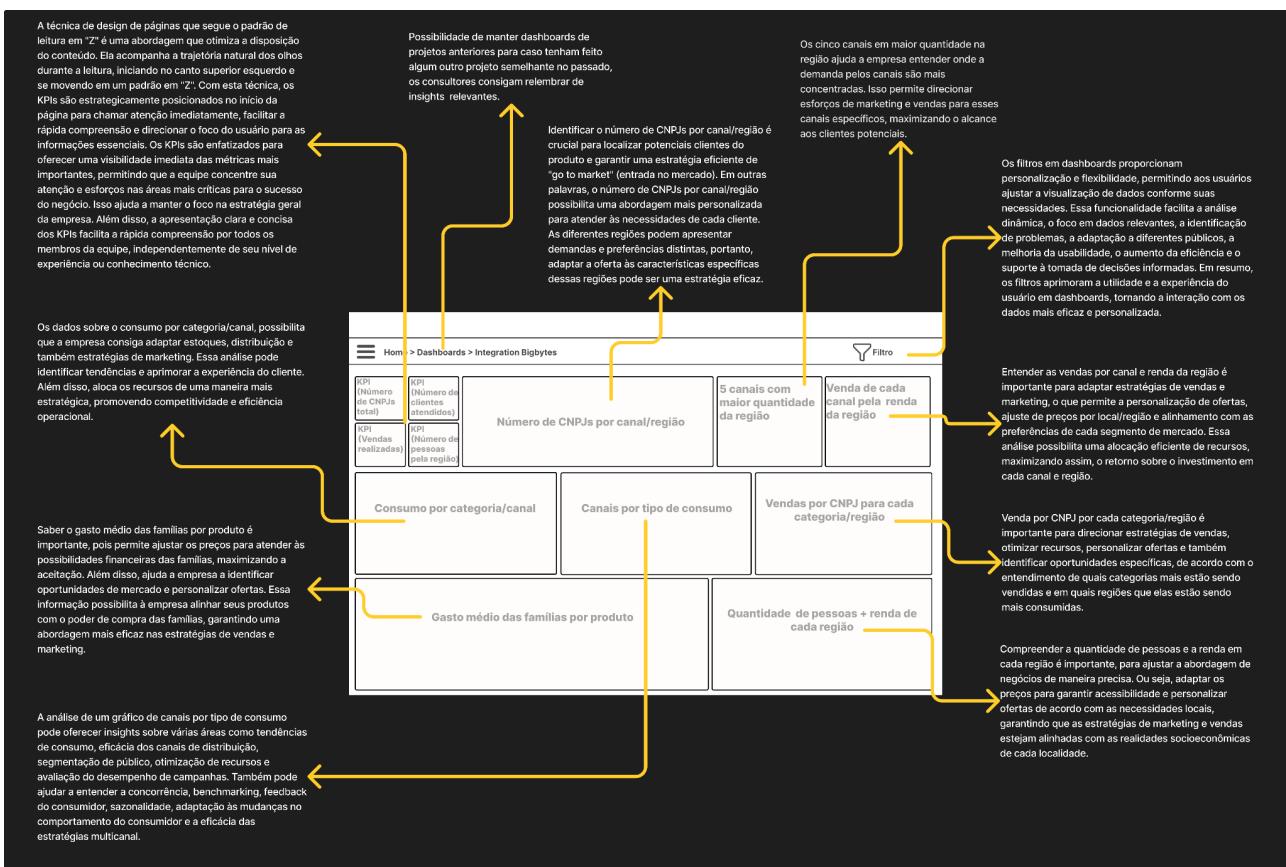


Imagen 19 - Descrição wireframe 1

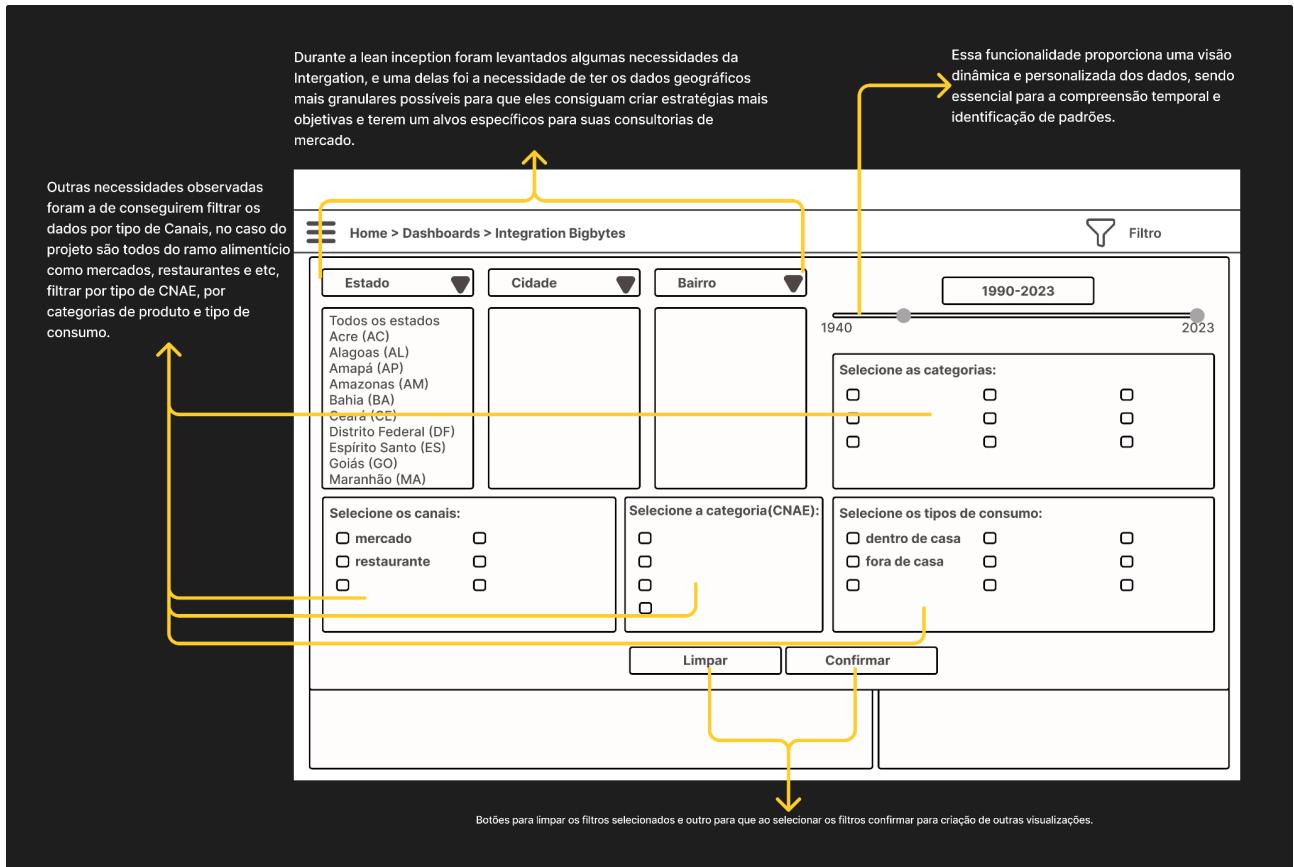


Imagen 20 - Descrição wireframe 2

Para melhor visualização, segue o link:

<https://www.figma.com/file/c6xWg8T5VXaamlURXYTpMT/wireframe-doc?type=whiteboard&node-id=0-1&t=wiPYqSI1WMNprcXz-0>

5. Arquitetura Macro

A arquitetura de dados é concebida para suportar um fluxo contínuo e eficiente de informações. Com base nos dados fornecidos pelo cliente, os requisitos identificados incluem:

- **Cubo de Informações:** Estruturado para acumular dados como o total de CNPJs, número de clientes atendidos, transações de vendas e potencial de mercado.
- **Automatização do Pipeline:** Configuração de um pipeline de dados automatizado para processamento e análise contínuos.
- **Data Lake:** Criação de um Data Lake estruturado para armazenar dados em grande escala.

- **Streaming de Dados:** Implementação de um fluxo de dados baseado em streaming para dados específicos, visando a redução do ruído informacional e a eficiência na transmissão de dados.
- **Preparação de Dados:** Desenvolvimento de um pipeline de dados dedicado à preparação e transformação de informações, alinhando-as aos objetivos do projeto.
- **Governança de Dados:** A API do cliente é projetada para não armazenar informações diretamente no cubo de dados, mantendo a governança e segurança das informações.
- **Análise Histórica:** Capacidade de analisar dados históricos ao longo de uma década.
- **Flexibilidade do Pipeline:** Construção e incremento facilitados do pipeline para acomodar novas fontes de dados e mudanças nos requisitos de negócios.
- **Visualização:** Disponibilização de ferramentas de visualização para interpretar informações diretamente dos dados armazenados.
- **Atualizações Anuais:** Estratégia para atualizar os dados anualmente conforme novas informações governamentais são disponibilizadas.
- **Segurança:** Implementação de autenticação de segurança robustas para proteger os dados em todas as fases do fluxo.

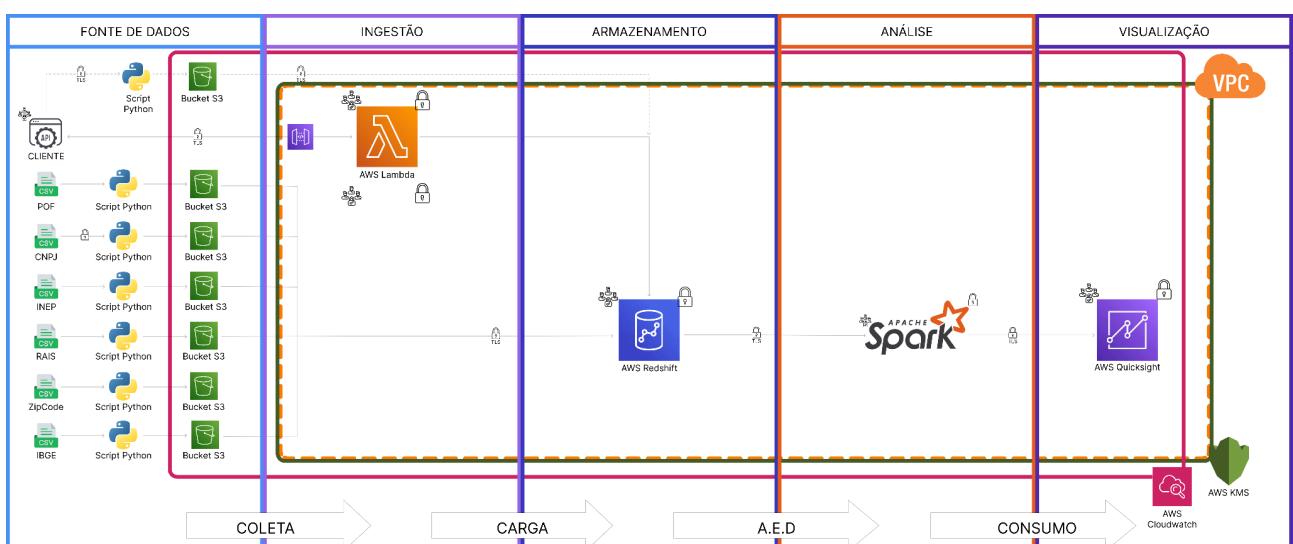


Imagen 21 - Descrição Arquitetura

A arquitetura projetada, ilustrada no diagrama anexo, proporciona um ciclo de vida completo dos dados, desde a coleta até a visualização, integrando uma variedade de serviços AWS para maximizar a eficiência e a escalabilidade.

5.1 Componentes da Arquitetura

5.1.1.API do cliente

Essa API atua como ponto de entrada para os dados. A "API do cliente" aceita dados enviados por clientes.

5.1.2.CSVs

São formatos de arquivo que contém dados tabulares. Os dados são melhores descritos na Análise Exploratória da seção abaixo.

5.1.3.Script Python

São aplicados para processar dados de várias fontes abertas, incluindo POF, INEP, Open DataSUS, MEC, ZipCode, além de CNPJs e também dos dados da API do cliente, preparando-os para o upload no S3. Além disso, um script específico é aplicado no Lambda para processar os dados mais recentes do cliente a cada determinado período de tempo.

5.1.4.Bucket S3

O bucket S3 é um serviço de armazenamento de objetos oferecido pela Amazon Web Services que oferece escalabilidade, disponibilidade de dados, segurança e performance. Ele organiza dados em 'buckets', que funcionam como contêineres básicos onde os dados são armazenados, e dentro dos quais são gerenciados através de chaves únicas (identificadores de objeto). Além disso, possui alta escalabilidade em termos de quantidade de dados armazenados e quantidade de solicitações por segundo, atendendo a necessidades de aplicações de todos os tamanhos. O Bucket S3 será utilizado para o armazenamento de Data Lake dos dados.

5.1.5.Lambda

O Lambda é um serviço de computação sem servidor que executa código em resposta a eventos e gerencia automaticamente os recursos de computação necessários. Ele é responsável por disparar funções de processamento de dados em resposta a eventos específicos, como uploads de dados da API para o RedShift. Além disso, escala

automaticamente ao processar cada evento individual, garantindo que a carga de trabalho seja gerenciada sem intervenção manual e executa códigos sem a necessidade de provisionar e gerenciar servidores, o que reduz a complexidade e o custo de operação.

5.1.6.Spark

Apache Spark é uma plataforma para processamento de dados em grande escala. Os dados do DynamoDB podem ser importados para o Spark para análises avançadas, processamento distribuído e aplicação de métodos ensamble.

5.1.7.VPC (Virtual Private Cloud)

Representa um ambiente isolado na AWS onde todos esses serviços e operações estão ocorrendo. Isso garante que os dados e serviços sejam seguros, isolados de outros recursos e facilmente gerenciáveis.

5.1.8.AWS CloudWatch

Implementado para o monitoramento contínuo dos logs e da telemetria, assegurando a disponibilidade e integridade dos dados.

5.1.9.AWS Redshift

O AWS Redshift é um serviço de armazenamento de dados rápido, escalável e totalmente gerenciado da Amazon Web Services, projetado especificamente para análise de grandes volumes de dados. A escolha do Redshift para um cubo de dados OLAP (Online Analytical Processing) é baseada em várias de suas características e capacidades que o tornam apropriado para essa finalidade:

- **Desempenho de Consulta:** Otimizado para executar consultas complexas rapidamente, permitindo análises de dados eficientes.
- **Escalabilidade:** Capacidade de expandir recursos conforme a necessidade de processamento e armazenamento de dados aumenta.
- **Gerenciamento Simplificado:** Automatiza tarefas operacionais, facilitando o gerenciamento de data warehouses.
- **Integração com Ferramentas de BI:** Conecta-se facilmente com ferramentas de Business Intelligence para visualização e análise de dados.

- **Custo-eficiência:** Oferece um modelo de pagamento flexível, adequado para orçamentos variados.
- **Segurança Robusta:** Inclui recursos avançados de segurança para proteger dados.
- **Conformidade Regulatória:** Adere a padrões de conformidade, essencial para setores regulados.
- **Análise Multidimensional:** Suporta a criação de cubos OLAP para uma análise de dados rica e multidimensional.
- **Insights Profundos:** Habilita as organizações a extrair insights valiosos dos seus dados para informar a tomada de decisão estratégica.

5.1.10.Segurança

Além disso, foram considerados aspectos de segurança, a fim de garantir a integridade e LGPD aos dados: 1.Criptografia de dados em trânsito: Como os dados de CNPJ e POF são de fontes governamentais, portanto públicas, não é necessário criptografá-los até chegarem no Dynamo DB, onde se juntam aos dados do cliente. A criptografia em trânsito é feita com o protocolo TLS. 2.Criptografia de dados em repouso: Todos os serviços que usam os dados do cliente devem ser configurados com opções de criptografia selecionadas, portanto AWS Lambda, Dynamo, Apache Spark e AWS Quicksight. 3.Grupos de acesso: Cada serviço deve ser acessado por pessoas autorizadas, ou seja, devem possuir a autenticação para seu devido acesso. Visto que tanto consultores de Tech&Digital, quanto de Marketing e Vendas se beneficiarão com a solução, cada um deve ter o acesso restrito à apenas sua necessidade na solução. 5.AWS Key Management Service: Para controlar os acessos, é utilizado o serviço da AWS.

5.2.Expурго dos dados

Para realizar a exclusão de dados do cliente armazenados no Amazon Redshift de forma automatizada, foi projetada uma função Lambda com um gatilho configurado para executar o expurgo periodicamente.

O propósito deste código é remover informações do cliente que não devem ser armazenadas no banco de dados para garantir a conformidade com políticas de proteção de dados.

A função Lambda está associada a uma função IAM para conceder as permissões necessárias de interação com o Redshift e é responsável por executar a exclusão dos dados específicos. O expurgo pode ser acionado por meio de um gatilho configurável, como por exemplo, a chegada de novos dados.

A implementação deste código de expurgo foi uma demanda do próprio cliente pela necessidade de garantir que informações confidenciais, relacionadas aos seus parceiros, fossem excluídas de maneira regular e automática, respeitando assim as normativas de privacidade e proteção de dados.

5.2. Análise Descritiva

A análise descritiva é uma abordagem fundamental na interpretação e compreensão de conjuntos de dados. Este método estatístico visa descrever e resumir as principais características dos dados, oferecendo uma visão clara e concisa das tendências, padrões e distribuições presentes. Por meio da análise descritiva, buscamos extrair insights valiosos que facilitam a interpretação e a comunicação efetiva dos resultados, contribuindo para uma compreensão mais profunda do fenômeno estudado.

5.2.1.Base dos [dados.org](#) - Base dos Dados

Organização sem fins lucrativos - GitHub - “Base dos Dados” é uma organização sem fins lucrativos que tem como missão universalizar o acesso a dados de qualidade para todos. Eles fornecem uma variedade de dados, incluindo estimativas populacionais, dados geográficos, informações sobre educação básica, inflação, Produto Interno Bruto (PIB).

5.2.2.IBGE Dados Abertos - Dados Abertos | IBGE

Governamentais - CSV - “Dados Abertos | IBGE” é uma iniciativa do Instituto Brasileiro de Geografia e Estatística (IBGE) que visa promover a transparência e a acessibilidade dos dados coletados pelo instituto. O IBGE é uma das principais fontes de dados estatísticos sobre o Brasil, fornecendo informações valiosas em diversas áreas, como demografia, economia, geografia.

5.2.3.POF (pesquisa orçamento familiar - POF 2017-2018 | IBGE

Governamentais - CSV - A Pesquisa de Orçamentos Familiares (POF) é um levantamento realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no Brasil. A POF tem como objetivo coletar informações detalhadas sobre os gastos e o padrão de vida das famílias brasileiras. O levantamento é conduzido a cada 5 anos, o que permite ao IBGE monitorar e analisar as mudanças nos hábitos de consumo e nas condições de vida da população.

5.2.3.1.Tabelas POF

aluguel_estimado: dados relacionados a domicílios e renda incluindo valores, formas de aquisição, meses de realização e informações sobre o rendimento bruto total mensal da unidade de consumo de várias unidades federativas do Brasil. Cada linha representa um domicílio e fornece detalhes sobre sua situação econômica. Os dados desta tabela proporcionam uma visão do potencial de consumo em diferentes canais e regiões, podendo contribuir para otimizar as estratégias do cliente no mercado alimentar e food service.

caderneta_coletiva: dados da pesquisa POF 2017-2018 no Brasil com informações sobre aquisições e despesas de domicílios de diferentes unidades da federação. Ela é útil

para análises que incluem os padrões de gastos e o orçamento familiar de diferentes regiões do Brasil. coletivas ou compartilhadas pelas famílias, como despesas ou investimentos em comum.

características_dieta: informações coletadas durante a POF de 2017-2018 no Brasil, onde cada linha corresponde a um domicílio. Cada domicílio possui diversos detalhes incluídos como: hábitos alimentares, estrato de pesquisa, situação do domicílio, localização geográfica e informações de renda e saúde. Essa tabela permite uma compreensão dos comportamentos e das condições de vida dos domicílios de diferentes regiões do Brasil por fornecer uma visão das características das famílias brasileiras.

condicoes_vida: dados de quatro domicílios na pesquisa POF 2017-2018. Ela possui informações como rendimentos familiares e padrão de vida de várias áreas, o que pode ser útil para entender o panorama socioeconômico da região. Além disso, os fatores de expansão são fornecidos para permitir análises e estimativas populacionais.

consumo_alimentar: informações sobre o consumo de alimentos, especificando os tipos de alimentos, quantidades consumidas, valores nutricionais e renda total da unidade de consumo. Essa tabela é fundamental para compreender as condições financeiras das famílias e ter insights relacionados ao padrão de consumo de determinada região.

despesa_coletiva: dados sobre despesas coletivas da POF que descrevem serviços ou aquisição de produtos, como o tipo de produto, quantidade e a forma de aquisição. Esses registros são essenciais para compreender tendências de consumo a partir de uma análise do comportamento de consumo dos domicílios.

despesa_individual: informações sobre despesas individuais, ou seja, gastos específicos de cada membro da família. Esses dados incluem detalhes como tipo de despesa, valor gasto, localidade do domicílio, entre outros. Essa tabela pode ser útil na tomada de decisões em área como marketing por permitir uma análise dos comportamentos de consumo e como as despesas se distribuem.

domicilio: informações sobre as características dos domicílios, como tamanho, localização e condições de moradia (materiais de construção, abastecimento de água, uso de energia e número de cômodos). Essa tabela é fundamental pois permite análises que impulsionam ações em áreas da saúde e de infraestrutura.

inventario: registros de inventário de bens duráveis de domicílios, incluindo informações como unidade federativa, tipo de situação e número do domicílio, tipo e quantidade de bem durável adquirido, entre outros. Esses dados permitem análises sobre padrões de consumo e correlações com a renda familiar e outras variáveis.

morador_quali_vida: informações sobre as características dos domicílios, incluindo dados sobre educação, sexo, composição familiar e renda disponível per capita. Essa tabela é útil para compreender padrões de consumo e suas tendências.

morador: informações relacionadas a renda e educação sobre os membros da família, incluindo o nível de instrução do morador e sua renda disponível. Esses dados são valiosos pois permitem uma análise sobre a relação o acesso à educação e distribuição de renda.

outros_rendimentos: informações sobre rendimentos recebidos por domicílios, incluindo unidade da federação, número do domicílio, código do informante, entre outros. Essa tabela pode ser útil para compreender as nuances da distribuição de renda em diferentes contextos sociais.

rendimento_trabalho: dados relacionados aos rendimentos obtidos por meio do trabalho, como salários e remunerações. Eles são essenciais para entender as condições financeiras de vida das famílias.

restricao_produtos_servicos_saude: informações sobre a restrição ou acesso a produtos e serviços de saúde pelas famílias. Eles incluem a situação do domicílio, estrato de amostragem, dados de renda mensal, entre outros. Essa tabela é essencial para compreender o comportamento financeiro relacionado a saúde.

servico_nao_monetario_pof2: informações sobre tipos e formas de despesas/aquisições não monetárias em domicílios. Essa tabela permite uma análise dos padrões de consumo e ajuda a entender o comportamento do consumidor.

servico_nao_monetario_pof4: informações: informações de despesas e aquisições não monetárias em domicílios, incluindo localização do domicílio, valores e detalhes sobre a despesa, entre outros. Esses dados são valiosos para entender como modelar o orçamento familiar em diferentes contextos sociais no Brasil.

5.2.4.RAIS e CAGED Microdados - Microdados RAIS e CAGED — Ministério do Trabalho e Emprego (www.gov.br)

Governamentais - CSV - RAIS (Relação Anual de Informações Sociais) é um sistema de informação criado pelo Ministério do Trabalho e Emprego (atualmente incorporado ao Ministério da Economia) que reúne dados das empresas e estabelecimentos empregadores no Brasil. Já o CAGED (Cadastro Geral de Empregados e Desempregados**)** é outro sistema relacionado ao mercado de trabalho brasileiro, também mantido pelo Ministério da Economia. Ele é utilizado para registrar as admissões e demissões de empregados sob o regime da Consolidação das Leis do Trabalho (CLT).

5.2.5.Receita Federal Dados Abertos - Dados Abertos — Receita Federal (www.gov.br)

Governamentais - CSV - Os "Dados Abertos" disponibilizados pela Receita Federal se referem a informações e conjuntos de dados que a Receita Federal do Brasil torna disponíveis ao público em geral para consulta e uso. Esses dados abertos geralmente incluem informações relacionadas a tributação, arrecadação, cadastros de contribuintes, normas fiscais e outros aspectos relacionados à administração fiscal e aduaneira no Brasil.

Governamentais - CSV - Dados Abertos MEC é um sistema de informação mantido pelo Ministério da Educação que oferece diversos dados sobre a educação no Brasil.

5.2.6.Dados Abertos - MEC - Página inicial (www.gov.br)

Governamentais - CSV - Dados Abertos INEP é um sistema do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira que reúne microdados educacionais detalhados.

5.2.7.Microdados — Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira | Inep (www.gov.br)

Governamentais - CSV - Open Data SUS é uma plataforma mantida pelo Ministério da Saúde que disponibiliza dados abertos relacionados à saúde no Brasil.

5.2.8.Bem vindo - OPENDATASUS (saude.gov.br)

Internacionais - CSV - From-to Zip Code to Lat-Long é uma base de dados que oferece informações sobre códigos postais de todos os países e suas respectivas coordenadas geográficas.

5.2.8.Zip and Postal Codes of All Countries | Database Hub (back4app.com)

CNPJs - CSV - Storage Account com as bases de CNPJ é um conjunto de dados que provêm de registros oficiais sobre empresas no Brasil, categorizados pelo Código Nacional de Atividade Econômica (CNAE). As bases foram divididas em quatro partes, cada uma representando um subconjunto distinto de CNAEs:

- CNPJs - Parte 1 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 2 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 3 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 4 (cnpjlake.blob.core.windows.net)

CNPJs - CSV - Estes dados são derivados de registros oficiais de empresas, especificamente relacionados aos CNPJs e categorizados pelo Código Nacional de Atividade Econômica (CNAE):

- CNPJs - Parte 1 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 2 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 3 (cnpjlake.blob.core.windows.net)
- CNPJs - Parte 4 (cnpjlake.blob.core.windows.net)

5.2.9.API - JSON - BASE DE VENDAS FICTÍCIA VIA ENDPOINT INTEGRATION

A API é uma base de dados simulada que representa as transações de vendas de um distribuidor fictício. Esta base é disponibilizada através de um endpoint específico e é formatada em JSON, uma linguagem leve de intercâmbio de dados de fácil leitura para humanos e máquinas. O conteúdo deste dataset é atualizado diariamente, incorporando novas vendas que são associadas a alguns dos CNPJs listados anteriormente, e se concentra nas categorias de produtos de interesse. Esta simulação serve para testes, análises e desenvolvimento sem comprometer dados reais ou sensíveis.

5.3.Análise Exploratória

A análise exploratória é uma etapa fundamental no processo de investigação e compreensão de conjuntos de dados. Nesta fase, são identificados padrões, tendências e insights iniciais que possam orientar investigações mais aprofundadas. Ao empregar técnicas estatísticas e gráficas, a análise exploratória proporciona uma visão inicial dos dados, auxiliando na formulação de perguntas mais específicas e na definição de abordagens subsequentes. Este processo é essencial para revelar características essenciais dos dados e orientar de maneira informada os passos seguintes na análise.

Desenvolvemos Jupyter notebooks onde criamos as análises exploratórias de cada uma das bases de dados. Para acessar clique no link:
<https://github.com/2023M8T4Inteli/grupo1/tree/main/src/Analise%20Exploratoria>

5.4.Pipeline de BigData na AWS

O pipeline de Big Data é um sistema complexo que abrange várias etapas, desde a coleta até a visualização de dados. Ele é projetado para processar, armazenar e analisar grandes volumes de informações, permitindo que as organizações tomem decisões informadas com base em dados.

Nesse contexto, a arquitetura proposta envolve diversos componentes e serviços da AWS (Amazon Web Services), cada um desempenhando um papel específico no fluxo de dados.

A primeira etapa, a fonte de dados, marca o ponto inicial do processo. Aqui, os dados são coletados de diversas fontes externas, como APIs, bancos de dados, arquivos CSV, e

outros. A variedade de origens é ampla, incluindo informações sobre clientes, transações, vendas e dados governamentais, entre outros. Nesta fase, os dados brutos são adquiridos e preparados para a próxima etapa.

O conjunto de dados referente aos CNPJs, armazenado em formato CSV na Storage Account, é composto por registros oficiais sobre empresas no Brasil, categorizados pelo Código Nacional de Atividade Econômica (CNAE). Essas bases foram estrategicamente divididas em quatro partes, cada uma representando um subconjunto distinto de CNAEs. Outra fonte essencial é a Pesquisa de Orçamentos Familiares (POF) do IBGE, apresentada em formato CSV. Realizada a cada cinco anos, essa pesquisa detalhada visa coletar informações sobre os gastos e o padrão de vida das famílias brasileiras, proporcionando uma visão abrangente das mudanças nos hábitos de consumo e nas condições de vida da população. Além disso, a base de dados simulada de vendas fictícias, disponibilizada através de uma API em formato JSON, oferece transações diárias de um distribuidor fictício, conectando-se aos CNPJs previamente mencionados. Por fim, a base "Zip and Postal Codes of All Countries" em formato CSV, proveniente do Database Hub, fornece informações cruciais sobre códigos postais de todos os países, acompanhados de suas coordenadas geográficas, enriquecendo ainda mais o panorama de dados disponíveis.

Na segunda fase, ingestão de dados, os dados coletados são processados e encaminhados para o pipeline de ingestão. Isso é realizado por meio de scripts Python e AWS Lambda. Os dados são armazenados nos buckets S3 da AWS (Amazon Simple Storage Service), garantindo escalabilidade e durabilidade para futuras análises.

No processo de ETL (Extração, Transformação e Carga), a ingestão de dados é uma etapa crítica que visa preparar e armazenar as informações de forma adequada. Esses dados, muitas vezes heterogêneos em formato e qualidade, podem conter redundâncias e inconsistências.

Após a coleta, entra em cena a transformação de dados, uma fase crucial de limpeza e organização. A limpeza envolve a remoção de duplicatas e valores nulos, enquanto a transformação inclui a normalização e estruturação dos dados para facilitar análises futuras. Essa etapa é essencial para garantir a integridade e qualidade dos dados.

Os dados limpos e transformados são, então, carregados no AWS S3. O formato escolhido é o CSV, devido à sua simplicidade e interoperabilidade. O S3 atua como um

repositório temporário e não estruturado, data lake, mantendo os dados acessíveis e prontos para transferência para o data warehouse.

O DataLake, como conceito centralizado de armazenamento, permite o processamento e uso de grandes volumes de dados em sua forma original. A escalabilidade e flexibilidade do AWS S3 tornam-no ideal para acomodar diversos tipos de dados, desde estruturados até não estruturados. Essas características, como escalabilidade, armazenamento de todos os dados e eliminação da gestão de servidores, tornam o serviço propício ao Data Lake.

Para as etapas de transformação e carga dos dados, a escolha recai sobre o uso de scripts Python. Essa decisão é respaldada pela facilidade de uso e manutenção, bibliotecas diretamente integradas às ferramentas AWS e Azure, além de portabilidade e interoperabilidade.

É relevante observar que, para dados provenientes da API do cliente, a ingestão ocorre de duas maneiras distintas. Dados mais antigos são adquiridos diretamente do API Gateway, sendo acionados pelo Lambda. Em contraste, dados diários são ingeridos por scripts Python e armazenados no bucket. O expurgo desses dados também é realizado pelo Lambda, conforme solicitação do parceiro, completando o ciclo de ingestão de forma eficiente e organizada.

No terceiro estágio, o armazenamento, os dados são transferidos para o AWS Redshift, um serviço de armazenamento de dados rápido e escalável. Cada tipo de dado (POF, CNPJ, CEP e API) possui seu próprio bucket S3 dedicado, organizando eficientemente as informações. O Redshift é otimizado para executar consultas complexas de forma ágil.

A arquitetura do Data Warehouse adota uma abordagem trifásica, composta pelas fases Worker, Raw e Trusted, para garantir a eficácia do processamento e armazenamento de dados. Na Fase Worker, que marca o início do pipeline de dados, ocorre a coleta inicial de dados brutos de diversas fontes, capturando-os em seu estado mais puro, sem tratamento prévio. Em seguida, na Fase Raw, os dados coletados são armazenados no Data Lake em seu formato original, proporcionando flexibilidade e atuando como um repositório temporário para grandes volumes de dados em diversas estruturas. A Fase Trusted, última etapa do pipeline, representa o momento em que os dados brutos passam por transformações e limpezas, sendo então carregados no Data Warehouse. Nesta fase, os dados estão estruturados e confiáveis, prontos para análises e relatórios.

No contexto do Data Warehouse, a utilização de views é fundamental. Essas views agem como representações virtuais de tabelas derivadas de uma ou mais fontes de dados, não armazenando dados fisicamente, mas facilitando consultas e manipulação de dados de maneira simplificada e eficiente.

A quarta fase, análise de dados, faz uso do Apache Spark para análises avançadas e processamento distribuído. Os dados são transformados e otimizados para consultas eficientes, enquanto análises exploratórias são realizadas para obter insights valiosos.

A utilização de modelos ensemble, como o Random Forest, no pipeline de Big Data é uma estratégia valiosa para aprimorar a precisão e o desempenho geral do sistema. O modelo ensemble combina as previsões de vários modelos mais simples, explorando a diversidade e complementaridade de cada um. Duas abordagens comuns são o "bagging", onde cada modelo é treinado em um conjunto de dados diferente e suas previsões são combinadas, e o "boosting", que treina modelos sequencialmente, concentrando-se em corrigir os erros dos modelos anteriores.

A implementação desses modelos ensemble no pipeline de Big Data oferece diversas vantagens. Primeiramente, busca-se equilibrar a variância e o viés dos modelos, proporcionando uma performance mais robusta. Além disso, esses métodos geralmente apresentam melhor desempenho, sendo mais robustos e complexos, com um custo computacional ligeiramente maior, mas com resultados superiores. Essa abordagem permite a adição de múltiplos vetores de preferência, enriquecendo a qualidade das análises ao observar comportamentos específicos em indivíduos com características semelhantes.

No contexto específico do pipeline, foram aplicados modelos como o algoritmo K-means, utilizado para agrupar dados em clusters, e o Random Forest, que cria várias árvores de decisão de maneira aleatória. A análise dos resultados inclui a visualização da distribuição de valores únicos em colunas específicas, a utilização do método do cotovelo para determinar o número ideal de clusters, a previsão de valores dos produtos pelo Random Forest e a avaliação dos resultados através do Mean Squared Error (Erro Quadrático Médio).

A avaliação e validação dos resultados foram conduzidas de acordo com o CRISP-DM, um modelo padrão para projetos de análise de dados, composto por fases como a Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Este modelo é iterativo, permitindo ajustes contínuos e refinamentos para garantir resultados desejados.

Na etapa de visualização, a ferramenta Metabase, conectada ao AWS Redshift, entra em cena. O Metabase permite a criação do infográfico, facilitando a análise de dados de maneira intuitiva.

Para atender às necessidades do parceiro, propõe-se a criação de um infográfico informativo e visualmente apelativo. A estratégia de visualização contempla diversas dimensões, cada uma respondendo a uma pergunta-chave sobre os consumidores e o mercado.

A primeira pergunta, "Quem são meus clientes consumidores?", é abordada considerando diferentes aspectos. A distribuição de renda dos consumidores será apresentada destacando as faixas de renda mais prevalentes. Além disso, entende-se os gastos médios dos consumidores em diferentes momentos de consumo e os horários mais comuns.

A segunda pergunta, "Quais canais quero estar presente?", é explorada através de um gráfico de barras, onde cada barra representa um canal de compra, e a altura indica a preferência relativa dos consumidores.

A terceira pergunta, "Em quais geografias quero atuar?", tem suas informações visualizadas em um mapa geográfico, destacando as regiões desejadas para atuação.

A quarta pergunta, sobre as dietas seguidas nos estados do país, é respondida através de um gráfico de mapa que destaca as dietas mais populares em diferentes regiões, padrões para representar os diferentes tipos de dietas.

A quinta pergunta, sobre os valores nutricionais dos alimentos consumidos no Brasil, será abordada por um gráfico de barras empilhadas. Cada barra representa uma unidade federativa, e as diferentes cores indicam os diversos componentes nutricionais.

Por fim, a sexta pergunta aborda a previsão de preço fornecida pelo modelo. Essa informação ressalta as tendências ou variações ao longo do tempo.

Já a última fase do pipeline aborda Segurança e Monitoramento. A segurança é mantida através do uso de TLS em várias partes do pipeline, e o AWS CloudWatch é empregado para monitorar métricas e logs, assegurando o desempenho e a confiabilidade do sistema.

No âmbito da segurança, foram minuciosamente considerados diversos aspectos para assegurar a integridade e atender aos requisitos da Lei Geral de Proteção de Dados (LGPD) no pipeline de Big Data. A criptografia desempenha um papel central nesse contexto, sendo aplicada de maneira estratégica.

Cada serviço é acessado somente por indivíduos autorizados, com autenticação necessária para garantir que apenas pessoas autorizadas tenham acesso aos recursos pertinentes. Tanto os consultores de Tech&Digital quanto os de Marketing e Vendas têm acesso restrito apenas às funcionalidades essenciais para suas atividades, fortalecendo a segurança global do sistema.

Essas medidas abrangentes, integradas em todas as fases do pipeline, não apenas preservam a confidencialidade dos dados sensíveis, mas também garantem a conformidade com regulamentações de privacidade. O compromisso com práticas robustas de segurança estabelece um ambiente confiável e protegido para o tratamento de dados críticos, consolidando a solução como uma base sólida para as operações da organização.

Em última análise, o pipeline de Big Data não apenas atende às necessidades atuais de análise de dados, mas também estabelece uma base sólida e adaptável para futuras evoluções. Sua abordagem iterativa e integrada, combinada com práticas sólidas de segurança, posiciona-o como uma solução robusta e confiável para impulsionar a tomada de decisões e insights estratégicos na organização.

6. Desenvolvimento do Cubo de Dados

Nesta fase, criamos uma estrutura que organiza e consolida dados relevantes para análises robustas. O cubo de dados proporciona uma visão abrangente e interativa, facilitando a extração de insights valiosos que impulsionam decisões informadas e estratégias eficazes.

6.1. Processo de ETL para Cubo de Dados OLAP

Essa seção descreve o processo de extração, transformação e carga (ETL) utilizado para consolidar dados de diversas fontes e alimentar um cubo de dados OLAP, utilizando as ferramentas da AWS para análise e visualização.

6.1.1. Visão geral

O objetivo deste pipeline de ETL é extrair dados de diversas fontes, transformá-los em um formato estruturado adequado para processamento analítico e carregá-los em uma solução de armazenamento de dados. O objetivo final é facilitar a análise de dados

complexos e multidimensionais para tarefas rápidas e interativas de inteligência de negócios.

6.2.Processo de ETL

O Processo de ETL (Extração, Transformação e Carga) é uma metodologia fundamental no domínio da integração de dados. Essa abordagem, essencial para a movimentação eficiente de dados entre diferentes sistemas, compreende três fases cruciais. Inicialmente, ocorre a Extração, onde dados são coletados de diversas fontes. Em seguida, a Transformação, proporcionando a limpeza, organização e conversão desses dados para um formato uniforme e utilizável. Por fim, a fase de Carga implica no carregamento dos dados transformados em um destino específico, preparando-os para análises e tomadas de decisão.

6.2.1.Extração de dados

O processo começa com a coleta de dados brutos de diversas fontes, incluindo APIs, bancos de dados e arquivos de texto. Esses dados são geralmente heterogêneos em formato e qualidade e podem conter redundâncias e inconsistências.

6.2.2.Transformação de dados

Após a coleta, os dados passam por uma fase crítica de limpeza e transformação. A limpeza envolve a remoção de duplicatas e valores nulos, enquanto a transformação inclui a normalização e estruturação dos dados para facilitar análises futuras. Esta etapa é essencial para garantir a integridade e a qualidade dos dados.

6.2.3.Carregamento no AWS S3

Os dados limpos e transformados são então carregados no serviço de armazenamento em nuvem AWS S3. O formato CSV é adotado por sua simplicidade e interoperabilidade. O S3 atua como um repositório temporário e não organizado, um data lake, que mantém os dados acessíveis e prontos para serem transferidos para o data warehouse.

Um Data Lake é um repositório centralizado que ingere e armazena grandes volumes de dados em sua forma original. Os dados podem ser processados e usados como base para uma variedade de necessidades analíticas. Devido à sua arquitetura aberta e

escalável, um Data Lake pode acomodar todos os tipos de dados de qualquer fonte, desde dados estruturados (tabelas de banco de dados, planilhas do Excel) até semiestruturados (arquivos XML, páginas da Web) e não estruturados (imagens, arquivos de áudio, tweets). Abaixo é possível entender as características que fazem esse serviço ser propício ao Data Lake:

- **Escalabilidade:** AWS S3 lida com a escala, agilidade e flexibilidade necessárias para combinar diferentes abordagens de dados e análises.
- **Armazenamento de todos os dados:** Como o Amazon S3 escala de forma econômica, praticamente sem limites, você pode armazenar todos os seus dados, de qualquer fonte, e desbloquear seu valor.
- **Eliminação da gestão de servidores:** Com as opções mais sem servidor para análise de dados na nuvem, os serviços de análise da AWS são fáceis de usar, administrar e gerenciar.

Para as etapas de transformação e carga dos dados, foi escolhido a utilização de script Python. Tal escolha é justificada pelos seguintes fatores:

1. Facilidade de uso e manutenção;
2. Bibliotecas ligadas diretamente às ferramentas AWS e Azure;
3. Portabilidade e Interoperabilidade.

Acesso aos scripts desenvolvidos: [Pasta - Script](#)

6.3. Armazenamento e análise de dados

OLAP (Online Analytical Processing) é uma tecnologia de banco de dados que foi otimizada para consulta e relatórios, em vez de processar transações. Um cubo OLAP é uma estrutura de dados montada de forma multidimensional, e que proporciona uma rápida análise de valores quantitativos ou medidas relacionadas com determinado assunto, sob diversas perspectivas diferentes. Os metadados do cubo são tipicamente criados a partir de um esquema de estrela ou esquema floco de neve de tabelas em um banco de dados relacional.

As fontes de dados utilizados a partir do Data Lake são:

- **IBGE Dados Abertos (CSV)**- “Dados Abertos | IBGE” é uma iniciativa do Instituto Brasileiro de Geografia e Estatística (IBGE) que visa promover a transparência e a acessibilidade dos dados coletados pelo instituto. O IBGE é uma das principais fontes de dados estatísticos sobre o Brasil, fornecendo informações valiosas em diversas áreas, como demografia, economia, geografia.
- **POF (pesquisa orçamento familiar - POF 2017-2018 | IBGE (CSV)** - A Pesquisa de Orçamentos Familiares (POF) é um levantamento realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no Brasil. A POF tem como objetivo coletar informações detalhadas sobre os gastos e o padrão de vida das famílias brasileiras. O levantamento é conduzido a cada 5 anos, o que permite ao IBGE monitorar e analisar as mudanças nos hábitos de consumo e nas condições de vida da população.
- **Receita Federal Dados Abertos (CSV)** - Os “Dados Abertos” disponibilizados pela Receita Federal se referem a informações e conjuntos de dados que a Receita Federal do Brasil torna disponíveis ao público em geral para consulta e uso. Esses dados abertos geralmente incluem informações relacionadas à tributação,

arrecadação, cadastros de contribuintes, normas fiscais e outros aspectos relacionados à administração fiscal e aduaneira no Brasil.

- **Dados Abertos MEC (CSV)**- Dados Abertos MEC é um sistema de informação mantido pelo Ministério da Educação que oferece diversos dados sobre a educação no Brasil.
- **Microdados Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (CSV)**- Dados Abertos INEP é um sistema do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira que reúne microdados educacionais detalhados.
- **OPENDATASUS (CSV)**- Open Data SUS é uma plataforma mantida pelo Ministério da Saúde que disponibiliza dados abertos relacionados à saúde no Brasil.
- **Zip and Postal Codes of All Countries | Database Hub (CSV)**- From-to Zip Code to Lat-Long é uma base de dados que oferece informações sobre códigos postais de todos os países e suas respectivas coordenadas geográficas.
- **CNPJs (CSV)**- Storage Account com as bases de CNPJ é um conjunto de dados que provêm de registros oficiais sobre empresas no Brasil, categorizados pelo Código Nacional de Atividade Econômica (CNAE). As bases foram divididas em quatro partes, cada uma representando um subconjunto distinto de CNAEs.
- **API (JSON) - BASE DE VENDAS FICTÍCIA VIA ENDPOINT INTEGRATION** é uma base de dados simulada que representa as transações de vendas de um distribuidor fictício. Esta base é disponibilizada através de um endpoint específico e é formatada em JSON, uma linguagem leve de intercâmbio de dados de fácil leitura para humanos e máquinas. O conteúdo deste dataset é atualizado diariamente, incorporando novas vendas que são associadas a alguns dos CNPJs listados anteriormente, e se concentra nas categorias de produtos de interesse.

Para armazenar e utilizar esses dados visando Business Intelligence, foi escolhido um Data Warehouse, que armazena dados que foram tratados e transformados com uma finalidade específica em mente, que podem ser usados para gerar insights. Em um banco de dados relacional, o esquema define as tabelas, campos, relacionamentos, visões, índices, pacotes, procedimentos, funções, filas, gatilhos, tipos, sequências, visões materializadas, sinônimos, enlaces de banco de dados, diretórios, entre outros elementos. Abaixo estão descritas as características que fazem o AWS Redshift ser a escolha ideal para essa função:

- **Alto desempenho:** O Redshift alcança alto desempenho usando paralelismo massivo, compressão de dados eficiente, otimização de consultas e distribuição.
- **Velocidade:** Quando se trata de carregar dados e consultá-los para análises e relatórios, o Redshift é extremamente rápido.
- **Segurança:** O Redshift tem recursos de segurança integrados, incluindo isolamento de rede, criptografia de descanso e autenticação IAM.
- **Suporte para novas funcionalidades SQL:** O Amazon Redshift suporta funcionalidades SQL, para simplificar a construção multidimensional do cubo e incorporar dados em rápida mudança.

6.3.1.Configuração do AWS Redshift

1. No Console da AWS, buscar por AWS Redshift

The screenshot shows the AWS Redshift landing page. At the top, there's a navigation bar with links like 'Painel do Corretor...', 'Corretor Online - P...', 'Bradesco', 'Sompo', 'HDI DIGITAL', 'Azul', 'Mapfre Connect', 'Portal Parceiros - To...', 'CEP', and 'Todos os marcadores'. Below the navigation is a search bar with 'Pesquisar' and a placeholder '[Alt+S]'. The main content area features the 'Amazon Redshift' logo and the tagline 'Fast, fully managed, petabyte-scale cloud data warehouse.'. A call-to-action button 'Try Redshift Serverless free trial' is visible. On the left, there's a sidebar with 'Análises' and 'Como funciona'. On the right, there's a 'Getting started' section with a 'Redshift Serverless overview' link. The bottom of the page includes standard AWS footer links: 'CloudShell', 'Comentários', '© 2023, Amazon Web Services, Inc. ou suas afiliadas.', 'Privacidade', 'Termos', and 'Preferências de cookies'.

Imagen 22 - AWS 1

2. Clique em 'Try Redshift Serveless' para abrir as configurações de criação do Redshift

The screenshot shows the 'Primeiros passos com o Amazon Redshift Serverless' configuration page. At the top, there's a blue header bar with the text 'Try new Amazon Redshift features in preview' and 'Create a workgroup with preview features. Production use of the workgroup is not supported. Use this workgroup for testing only.' There's also a 'Create preview workgroup' button. Below the header, the page title is 'Primeiros passos com o Amazon Redshift Serverless' with an 'Informações' link. A sub-section titled 'Configuração' contains two options: 'Usar configurações padrão' (selected) and 'Personalizar configurações'. The 'Namespaces' section follows, with an 'Informações' link. The bottom of the page includes standard AWS footer links: 'CloudShell', 'Comentários', '© 2023, Amazon Web Services, Inc. ou suas afiliadas.', 'Privacidade', 'Termos', and 'Preferências de cookies'.

Imagen 23 - AWS 2

3. Rolando a tela em 'Funções do IAM associadas', clique em 'Gerenciar funções do IAM' e logo em seguida em 'Criar função'.

The screenshot shows the AWS Lambda console with the following details:

- Header:** Shows the AWS logo, navigation menu, search bar, and account information ("Norte da Virgínia" and "dmsophia").
- Banner:** A blue banner at the top says "Try new Amazon Redshift features in preview" and "Create a workgroup with preview features. Production use of the workgroup is not supported. Use this workgroup for testing only." It also has a "Create preview workgroup" button.
- Page Content:**
 - A breadcrumb trail: "Amazon Redshift Serverless > Primeiros passos com o Amazon Redshift Serverless".
 - A main title: "Primeiros passos com o Amazon Redshift Serverless" with a "Informações" link.
 - A text block: "Para começar a usar o Amazon Redshift Serverless, configure seu data warehouse sem servidor e crie um banco de dados. Você receberá USD 300,00 de crédito para o uso do Redshift Serverless nesta conta."
 - A configuration section with two options:
 - Usar configurações padrão: "As configurações padrão foram definidas para ajudar você a começar. É possível alterá-las a qualquer momento mais tarde."
 - Personalizar configurações: "Personalize suas configurações de acordo com suas necessidades específicas."
 - A namespace section with a "Namespace" and "Informações" link.
- Footer:** Includes links for "CloudShell", "Comentários", and copyright information ("© 2023, Amazon Web Services, Inc. ou suas afiliadas."), as well as links for "Privacidade", "Termos", and "Preferências de cookies".

Imagen 24 - AWS 3

4. Selecione 'Sem buckets do S3 adicionais' e crie a função.

The screenshot shows the "Criar a função do IAM padrão" (Create a standard IAM function) dialog box:

- Left Panel:** Shows "Funções do IAM associadas" (Associated IAM functions) with a note: "Crie, associe ou remova uma função do IAM e defini-la como padrão." Below it is a search bar "Pesquisar função do IAM associada" and a "Funções do IAM" button.
- Right Panel:** The main dialog has the title "Criar a função do IAM padrão".
 - Information Block:** Describes how to associate an IAM profile for Redshift Serverless to execute LOAD and UNLOAD commands. It mentions the "AmazonRedshiftAllCommandsFullAccess" policy.
 - Bucket Selection:** A section titled "Especificar um bucket do S3 para a função do IAM que será acessada" (Specify an S3 bucket for the IAM function to access). It includes:
 - Sem buckets do S3 adicionais: "Crie a função do IAM sem especificar os buckets do S3."
 - Qualquer bucket do S3: "Permitir que os usuários que tenham acesso aos dados do Redshift Serverless também acessem qualquer bucket do S3 e seu conteúdo na conta da AWS."
 - Buttons:** "Cancelar" (Cancel) and "Criar função do IAM como padrão" (Create IAM function as default).

Imagen 25 - AWS 4

5. Role a tela até o fim e crie o Redshift Serverless. Aguarde.

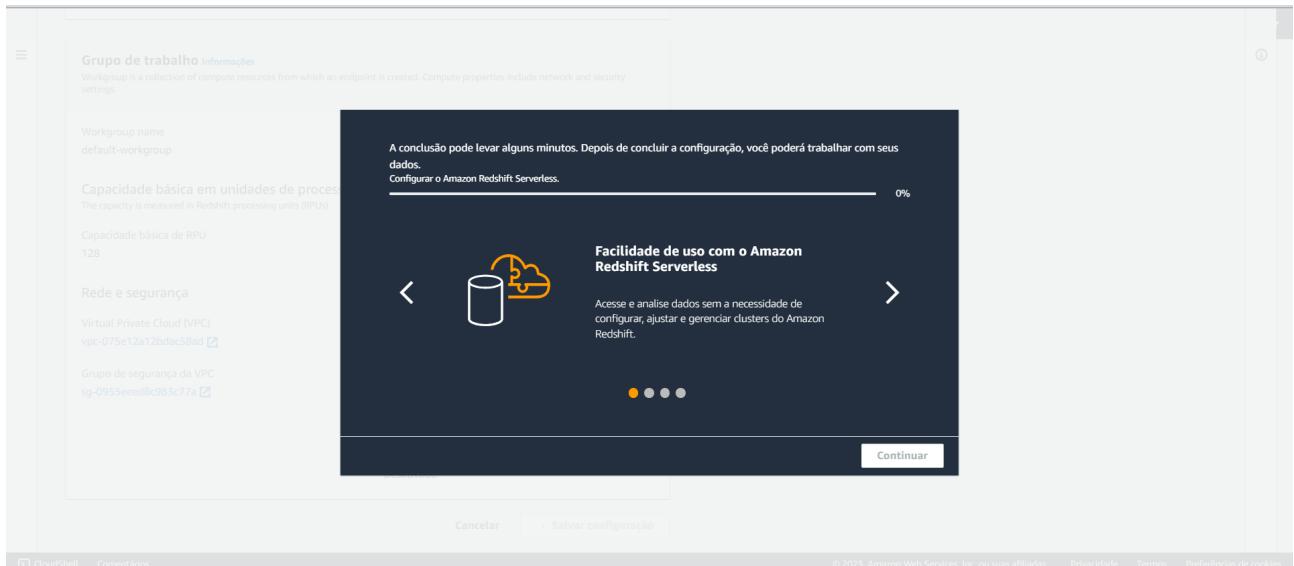


Imagen 26 - AWS 5

6. Após a conclusão, clique em continue. O Painel do Serveless será aberto, então selecione 'Consultar dados'.

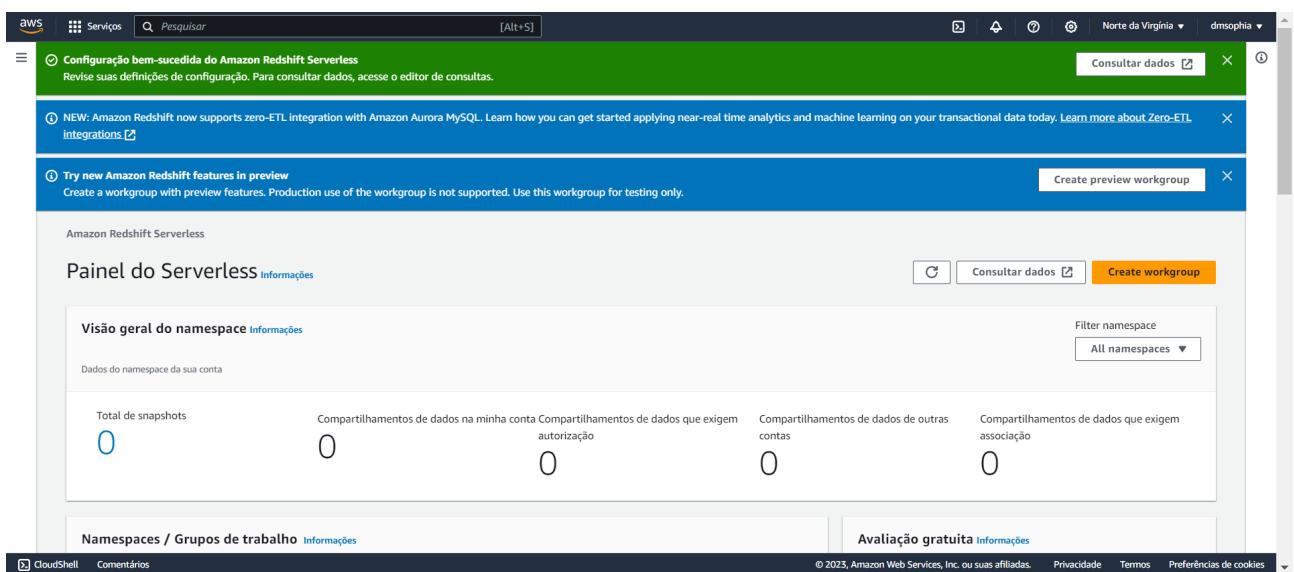


Imagen 27 - AWS 6

7. Em 'Consultar dados' essa será a página default.

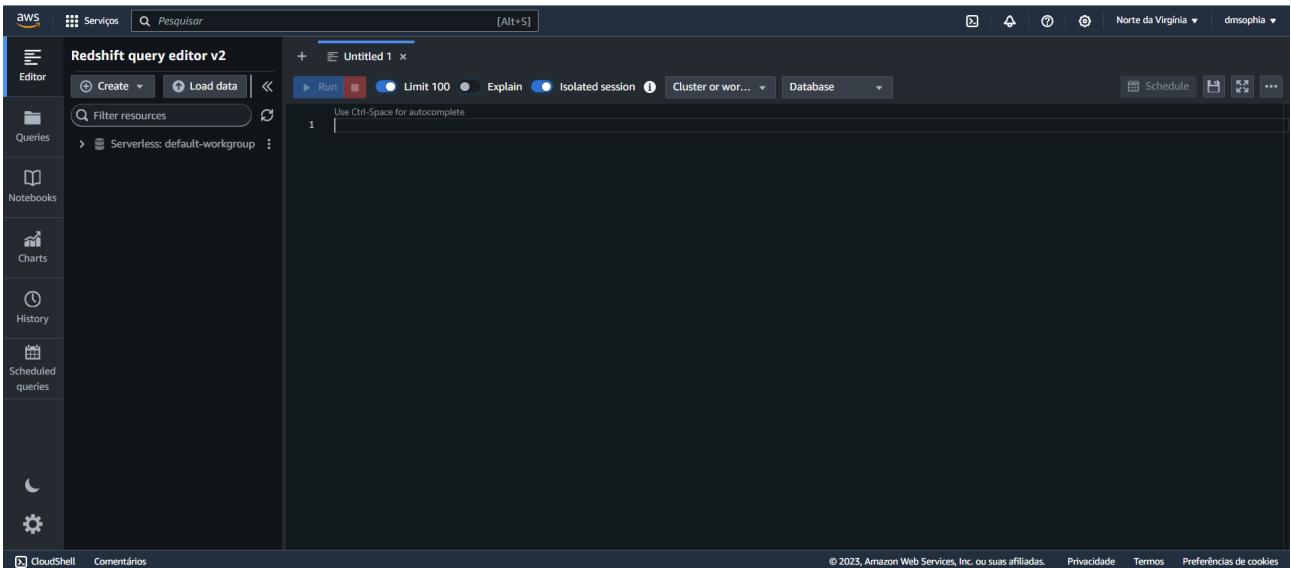


Imagen 28 - AWS 7

8. Clique em 'Load data' para carregar os arquivos desejados do Data Lake. Lembre-se que o Data Lake são os buckets do S3. Crie a conexão com o 'Federated User'.

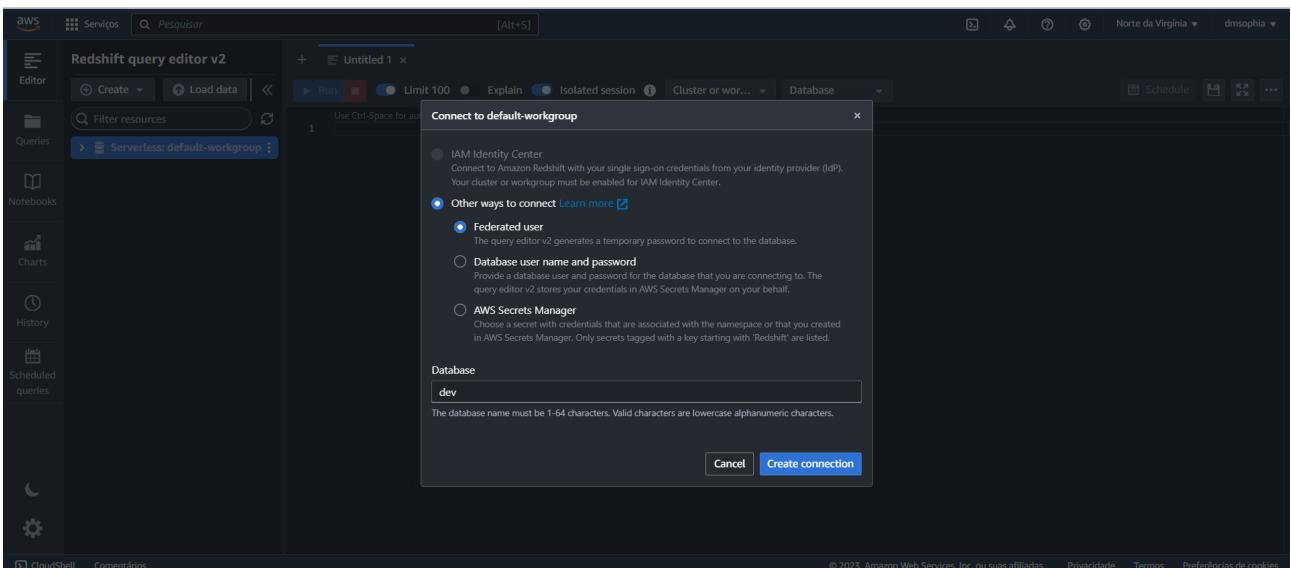


Imagen 29 - AWS 8

9. Carregue seus dados colocando a URI do seu bucket do S3 e a região dele.

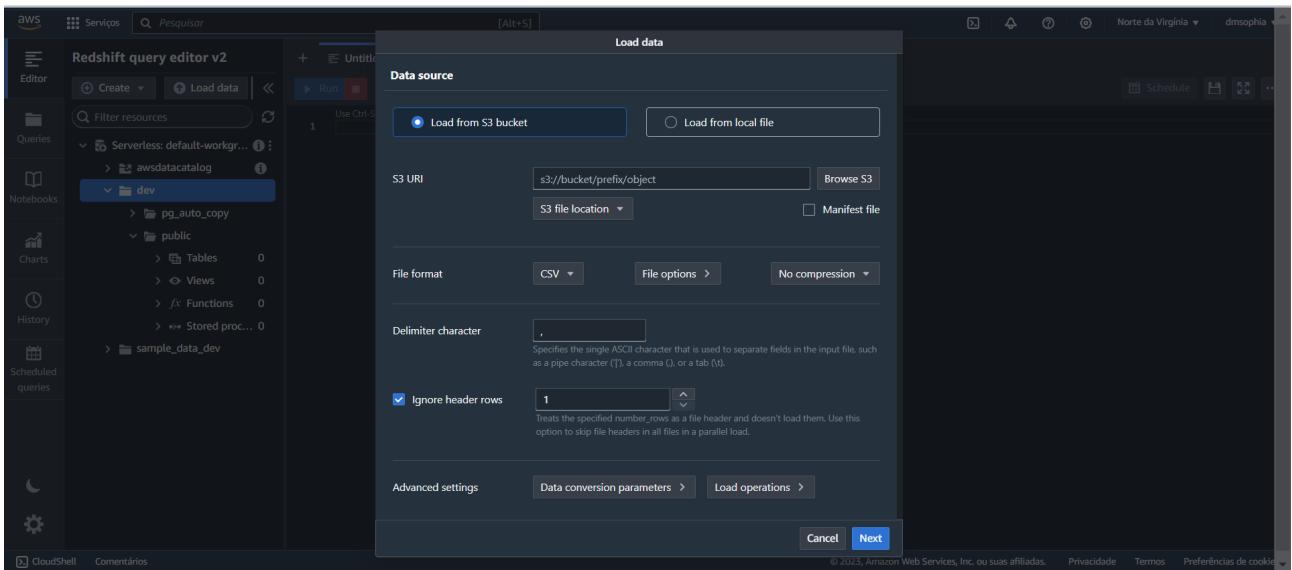


Imagen 30 - AWS 9

10. Agora faça a navegação: Serverless -> dev -> public. Em 'tables', será possível ver todas as suas tabelas carregadas dos buckets do S3 (realizado no passo anterior).

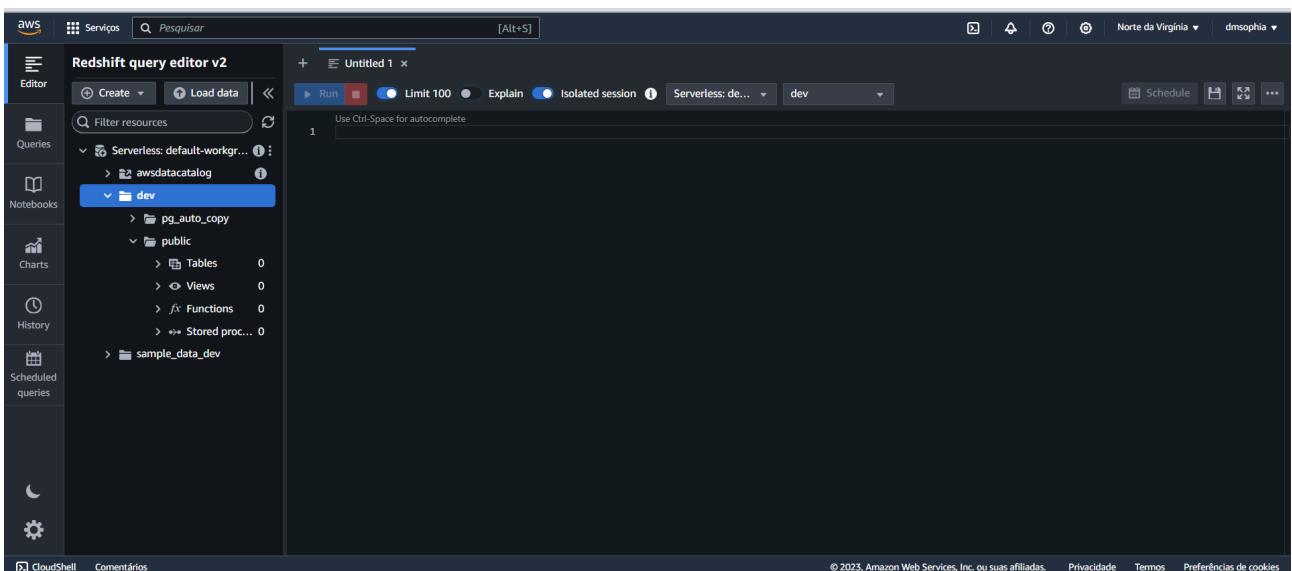


Imagen 31 - AWS 10

11. Ao lado direito, no editor, coloque suas manipulações SQL desejadas para tirar os insights das suas tabelas. As chamadas VIEWS, você poderá juntar dados das tabelas, criando outras, para então extrair insights.

The screenshot shows the AWS Redshift query editor interface. On the left, there's a sidebar with various navigation options like Editor, Queries, Notebooks, Charts, History, and Scheduled queries. The main area displays a SQL query in a tab titled 'Untitled 1'. The query is a CREATE VIEW statement for 'consumo_individual' with numerous columns listed. At the top of the editor, there are buttons for Run, Limit 100, Explain, Isolated session, and Serverless: dev... The status bar at the bottom indicates the session is 'dev'.

```

1 CREATE VIEW consumo_individual AS
2   SELECT
3     pof_despesaindividual.seq,
4     pof_despesaindividual.UF,
5     pof_despesaindividual.renda_total,
6     pof_despesaindividual.num_ur,
7     pof_despesaindividual.v9001 AS tipo_despesa,
8     pof_despesaindividual.v8800 AS valor_despesa,
9     pof_despesaindividual.v9004 AS local_aquisicao,
10    pof_consumoalimentar.v9015 AS horario_consumo,
11    pof_consumoalimentar.v9016 AS forma_preparacao,
12    pof_consumoalimentar.v9017 AS local_consumo,
13    pof_consumoalimentar.v9018 AS local_refeicao,
14    pof_consumoalimentar.DIA_SEMANA,
15    pof_consumoalimentar.DIA_ATIPICO
16   FROM
17   pof_despesaindividual
18   JOIN
19   pof_consumoalimentar
20   ON
21     pof_despesaindividual.seq = pof_consumoalimentar.seq;

```

Imagen 32 - AWS 11

6.3.2.Estrutura do Data Warehouse: Arquitetura de Três Fases

O Data Warehouse na arquitetura de dados adota uma estrutura trifásica, compreendendo as fases de Worker, Raw e Trusted, para assegurar a eficácia do processamento e armazenamento de dados.

- Fase Worker:** É o ponto de partida do pipeline de dados, onde ocorre a coleta inicial de dados brutos de fontes variadas. Aqui, os dados são capturados em seu estado mais puro, sem qualquer tratamento prévio.
- Fase Raw:** Nesta etapa, os dados coletados são armazenados no Data Lake em seu formato original. Esta fase serve como um repositório temporário e flexível, permitindo o armazenamento de grandes volumes de dados em diversas estruturas.
- Fase Trusted:** Representa a etapa final do pipeline. Aqui, os dados brutos passam por um processo de transformação e limpeza, sendo posteriormente carregados no Data Warehouse. Nesta fase, os dados estão prontos para serem utilizados em análises e relatórios, estando estruturados e confiáveis.

6.3.2.1.Utilização de Views no Data Warehouse

As views no Data Warehouse desempenham um papel crucial, agindo como representações virtuais de tabelas derivadas de uma ou mais fontes de dados. Estas views não armazenam dados fisicamente, mas facilitam a consulta e manipulação dos dados de forma simplificada e eficiente.

- **View "consumo_domicilio":** Essa view é criada pela junção das tabelas "pof_domicilio" e "pof_morador", utilizando a coluna "num_dom" como chave de junção. Ela consolida informações relevantes dos domicílios e dos moradores, como tipo de domicílio, presença de água canalizada, rendimento dos moradores, e outros dados socioeconômicos.
- **View "consumo_individual":** Formada pela junção das tabelas "pof_despesaindividual" e "pof_consumoalimentar", esta view oferece uma visão detalhada sobre os padrões de consumo e despesas individuais. Através dela, é possível analisar informações como o valor da despesa, tipo de produto consumido, local e horário de consumo, proporcionando insights valiosos sobre os hábitos de consumo.

Estas views são essenciais para proporcionar uma análise mais profunda e segmentada dos dados, permitindo que usuários e analistas acessem informações específicas de forma rápida e eficiente, sem a necessidade de consultar as tabelas de dados originais em sua totalidade. Os códigos para as views estão disponíveis em

6.4. Aspectos de segurança, privacidade e conformidade

A segurança de dados está mais focada em proteger a informação de ataques cibernéticos e violações. Já a privacidade trabalha a parte de como essa informação é coletada, compartilhada e utilizada. A arquitetura de dados é fundamental para organizar, padronizar e gerenciar informações valiosas para uma empresa. Afinal, ela é responsável por garantir a qualidade, segurança e integridade dos dados. Bem como disponibilizá-los de forma clara e eficiente a fim de apoiar a tomada de decisões estratégicas.

A privacidade de dados e o impacto na arquitetura são importantes para a conformidade com a LGPD (Lei Geral de Proteção de Dados), que pode exigir adequações nas estruturas físicas e arquitetônicas das empresas, especialmente diante da possibilidade de acessos indevidos por parte de terceiros. A LGPD possui amplo alcance, fazendo com que as empresas, independentemente do ramo de atuação, começam a estruturar uma linha de planejamento sobre a forma de proteção ao tratamento dos dados pessoais de seus clientes e colaboradores, a fim de evitar infrações, que certamente irão culminar em prejuízos financeiros e danos na reputação.

Na fase de coleta, os dados são adquiridos de várias fontes. A segurança envolve a autenticação para garantir que apenas fontes confiáveis possam fornecer dados. Além disso, os dados são criptografados durante a transmissão para proteger contra a interceptação.

Durante a ingestão, os dados são trazidos para o sistema. Novamente, a autenticação e a criptografia são usadas para garantir que apenas dados autorizados sejam ingeridos. Além disso, os dados podem ser validados neste ponto para garantir que eles atendam a certos padrões de qualidade e consistência.

Na fase de processamento, os dados são transformados e enriquecidos. A segurança envolve o controle de acesso para garantir que apenas usuários autorizados possam realizar operações de processamento.

Durante o armazenamento, os dados são mantidos para uso futuro. A segurança é relacionada à criptografia de dados em repouso para proteger contra acesso não autorizado. Além disso, os backups regulares são realizados para proteger contra perda de dados.

6.5.Monitoramento e gerenciamento do processo de ETL

O Amazon CloudWatch é um serviço de monitoramento e gerenciamento que tem como função fornecer dados práticos para recursos de aplicativos e infraestruturas locais, híbridos e da AWS. Com ele é possível coletar métricas e operações na forma de logs, propiciar visualização unificada dos recursos e resolver problemas. O Amazon CloudWatch coleta e visualiza logs, métricas e dados de eventos em tempo real em painéis automatizados para otimizar sua infraestrutura e manutenção da aplicação. Garante-se, então, o monitoramento, para pleno funcionamento dos scripts Python e a arquitetura na cloud.

6.6. VIEWS

As "Views" no Amazon Redshift representam uma ferramenta poderosa para simplificar a análise de dados. Ao criar uma "View", é possível encapsular consultas complexas em uma estrutura mais fácil de compreender, proporcionando uma visão organizada e simplificada dos dados armazenados no Redshift. Essa funcionalidade não apenas aumenta a eficiência na manipulação de informações, mas também oferece uma abordagem mais intuitiva para explorar e extrair insights valiosos dos conjuntos de dados. Nesta breve introdução, exploraremos como as "Views" no Amazon Redshift se tornam uma peça fundamental para otimizar a experiência analítica.

6.6.1. Consumo Individual

Tabela (consumo_individual)

Junção das tabelas de 'Despesa Individual (POF)' e 'Consumo Alimentar', com o intuito de visualizar como as despesas estão relacionadas ao consumo dos indivíduos.

6.6.2. Características do Consumo

Tabela (caracteristica_consumo)

Junção das tabelas 'Consumo Alimentar (POF)' e 'Características da Dieta (POF)' para entender os padrões de consumo e perfil dos consumidores, ou seja, como as dietas impactam no seu consumo, por região (estado).

6.6.3. CNPJs e Sales

Tabela (cnpj_union_view)

Junção das 4 tabelas de 'CNPJs' e 'Sales' da API do cliente para obter insights relacionados às vendas do cliente nas Unidades Federativas do país.

6.6.4. CNPJs, Sales e Category

Tabela (cnpj_union_view_category)

Junção das 4 tabelas de 'CNPJs', 'Sales' da API do cliente e 'Category', também da API do cliente, para entender, além das vendas, as categorias em granularidade de Unidades Federativas do país em relação apenas ao que se refere ao cliente.

7. Curadoria dos dados

O presente documento é responsável por fornecer os insumos necessários à justificativa de escolha de base de dados utilizados para as posteriores análises no cubo de dados. Foram selecionados até 15 GB de dados que auxiliam no principal objetivo: gerar insights sobre potencial de consumo por categoria, canal e região.

7.1. POF (Pesquisa de Orçamento Familiar)

A POF é uma pesquisa que detalha informações sobre a composição dos gastos das famílias em todo o Brasil. Esta pesquisa é crítica para o projeto, porque pode fornecer insights sobre o comportamento do consumidor, preferências e padrões de gasto dentro do setor alimentício, ou seja, o perfil do consumidor. Ao analisar os dados da POF, é possível entender quais tipos de produtos alimentícios são mais populares entre diferentes grupos demográficos, quanto eles estão dispostos a gastar e como esses padrões variam em diferentes regiões. Isso ajudará na segmentação do mercado e no direcionamento mais eficaz de grupos específicos de consumidores.

7.2. CNPJ

A base de dados do CNPJ é uma rica fonte de informações sobre as empresas no Brasil. Para os clientes da consultoria, esta base de dados é essencial porque inclui informações sobre a localização, tamanho e tipo de cada negócio registrado. É possível usá-la para identificar concorrentes no setor alimentício, analisar a saturação do mercado em diferentes regiões e avaliar o potencial para novos negócios. Ao saber quantos e quais tipos de empresas do setor alimentício estão operando em uma região, é possível aconselhar melhor os clientes sobre onde podem existir lacunas no mercado ou áreas de alta concorrência.

7.3. CEP

A base de dados de CEPs (Códigos de Endereçamento Postal) do Brasil é extremamente relevante para o projeto de análise do potencial de consumo para o setor alimentício. Aqui estão os motivos:

- 1. Geo-Referenciamento e Segmentação Regional :** CEPs permitem a localização precisa de estabelecimentos e consumidores. Isso é crucial para entender a distribuição geográfica dos clientes potenciais e para a segmentação de mercado

baseada em localização. É possível identificar áreas com alta concentração de bares, restaurantes, supermercados ou quaisquer outros estabelecimentos e, assim, cruzar essas informações com dados de consumo e demográficos para identificar regiões com alto potencial de consumo.

2. **Logística e Distribuição** : A análise de CEPs pode ajudar a otimizar a logística de distribuição dos clientes. Entender onde estão localizados os estabelecimentos permite planejar rotas de entrega mais eficientes e reduzir custos operacionais. Isso é particularmente importante para empresas de alimentos e bebidas que dependem de entregas rápidas e frescor dos produtos.
3. **Análise de Mercado e Expansão** : Ao combinar dados de CEP com informações de consumo e dados demográficos, é possível identificar áreas não atendidas ou com demanda insatisfeita. Isso pode direcionar estratégias de expansão para os clientes, indicando onde novos estabelecimentos poderiam ter sucesso ou onde poderiam aumentar sua participação no mercado.
4. **Planejamento Urbano e Demanda Futura** : Os CEPs também são úteis para analisar tendências de crescimento urbano e desenvolvimento de novas áreas. Isso pode antecipar a demanda futura e ajudar os clientes a planejar estrategicamente a abertura de novos negócios ou a expansão de negócios existentes.

8. Modelos Ensemble

Modelo ensemble é uma técnica em aprendizado de máquina que combina as previsões de vários modelos mais simples para melhorar a precisão e o desempenho geral do sistema. A ideia por trás dos modelos Ensemble é aproveitar a diversidade e a complementaridade de vários modelos individuais para obter um desempenho mais robusto e generalizável.

Existem diferentes abordagens para implementar modelos ensemble, mas duas das mais comuns são o "bagging" e o "boosting".

- **Bagging (Bootstrap Aggregating)**: No bagging, cada modelo é treinado em um conjunto de dados diferente, criado por meio de amostragem com reposição (bootstrap) do conjunto de dados original. Em seguida, as previsões de cada

modelos são combinadas por votação (no caso de classificação) ou por média (no caso de regressão) para produzir a previsão final.

- **Boosting:** No boosting, os modelos são treinados sequencialmente, e cada novo modelo se concentra em corrigir os erros cometidos pelos modelos anteriores.

A utilização de um modelo ensemble em um pipeline de big data é benéfica por várias razões. Primeiro, os métodos ensemble são uma tentativa de encontrar um equilíbrio entre variância e viés. Alguns modelos, por suas peculiaridades matemáticas, possuem maior variância ou maior potencial de overfitting e obter um viés. Segundo, os métodos ensemble geralmente têm uma performance melhor. Eles são mais robustos e complexos, envolvem mais operações, com um custo computacional um pouco maior, mas que, geralmente, têm uma performance melhor. Terceiro, eles permitem a adição de uma multiplicidade de novos vetores de preferência, complementando e enriquecendo a qualidade das análises devido à observação de comportamentos específicos em indivíduos com características similares

8.1 Algoritmo K-means

O algoritmo K-means é um método de aprendizado não supervisionado que é usado para agrupar dados em clusters. Primeiro, é preciso definir um 'K', correspondente ao número de agrupamentos que se deseja formar. Em seguida, define-se, de forma aleatória, um centróide para cada cluster. O centróide é basicamente o centro do cluster. O próximo passo é calcular, para cada ponto, a distância até o centróide mais próximo. Cada ponto será atribuído ao cluster do centróide mais próximo. Agora, reposiciona-se o centróide. A nova posição do centróide deve ser a média da posição de todos os pontos do cluster. Os dois últimos passos são repetidos, iterativamente, até que os centróides não mudem mais de posição.

Inicialmente, a fim de observar a relação entre os dados do data frame, utiliza-se o comando `'feat_categorical_nunique.hist()'` para criar um histograma dos valores únicos na coluna. Este é um método para visualizar a distribuição de valores únicos na coluna.

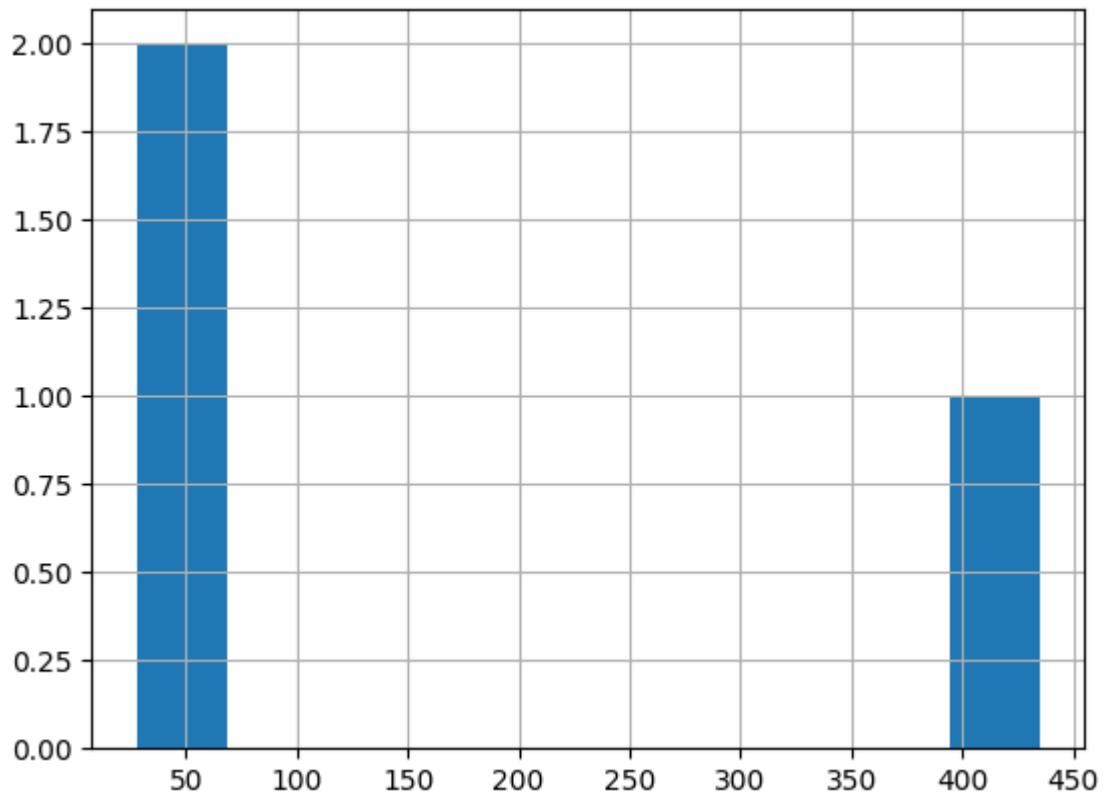


Imagen 33 - K-means

Em seguida, utiliza-se o método cotovelo para determinar o número ideal de clusters em um conjunto de dados. Ao plotar o gráfico da inércia em função do número de clusters, o eixo x representa o número de clusters e o eixo Y representa a inércia. O ponto onde a inércia começa a diminuir de forma linear é considerado o número ideal de clusters.

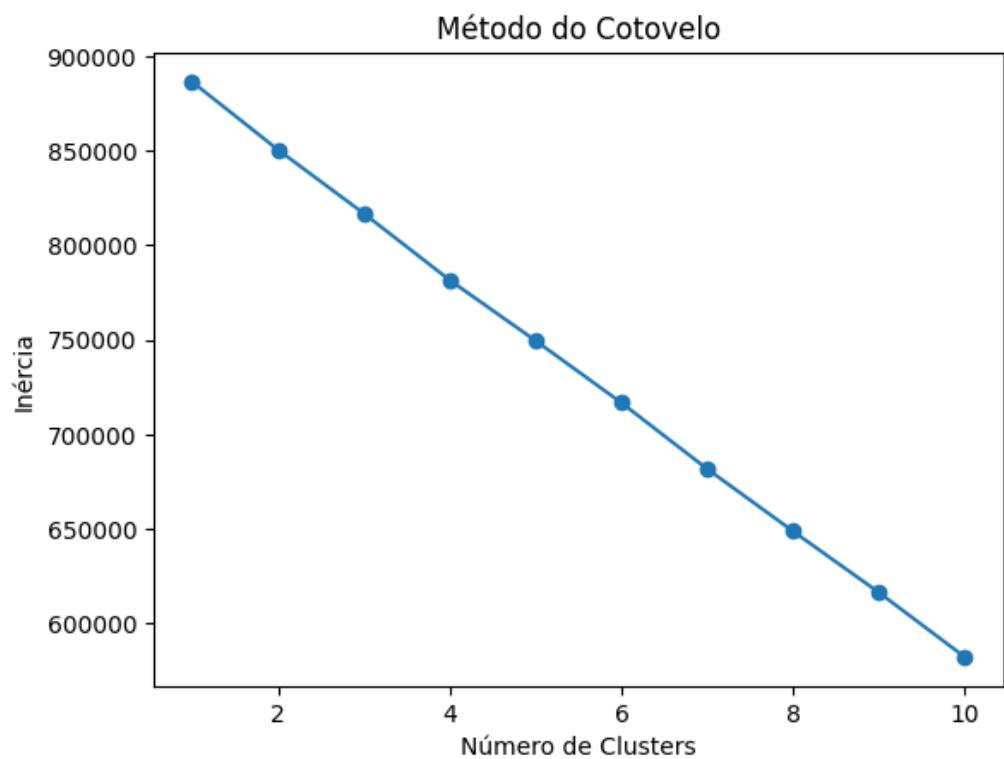


Imagen 34 - K-means 1

A partir de tal análise, é possível compreender que alguns estados do país possuem tendências no valor das suas compras. A escolha da quantidade de clusters ($k = 5$), é justificável ao selecionar o entendimento de região de compra.

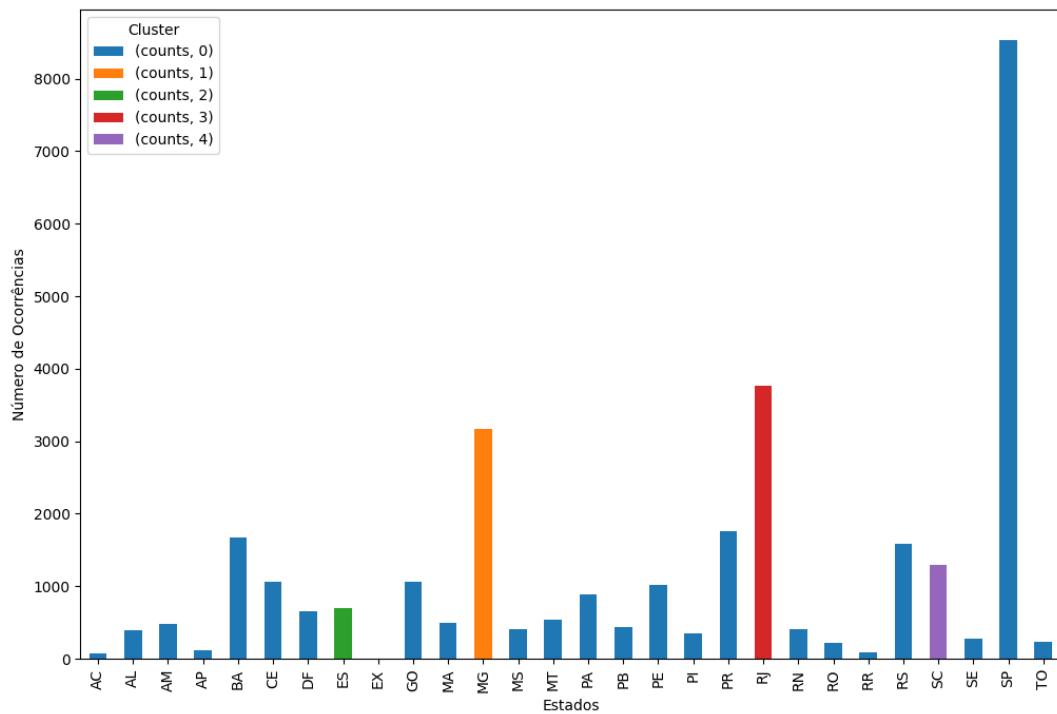


Imagen 35 - K-means 2

8.2 Modelo Random Forest

A Floresta Aleatória é um método de aprendizado conjunto que faz parte dos métodos ensemble. Ele cria várias árvores de decisão de maneira aleatória. Cada árvore é utilizada na escolha do resultado final, em uma espécie de votação.

As árvores de decisão, ou Decision Trees, estabelecem regras para tomada de decisão. O algoritmo cria uma estrutura similar a um fluxograma, com “nós” onde uma condição é verificada, e se atendida o fluxo segue por um ramo, caso contrário, por outro, sempre levando ao próximo nó, até a finalização da árvore.

A abordagem do Random Forest é ao mesmo tempo mais precisa e mais robusta para mudanças nas variáveis preditoras do que uma única árvore de classificação ou

regressão. Portanto, o Random Forest é uma forma de combinar vários modelos de machine learning em um único resultado.

O objetivo do modelo utilizado, alinhado ao projeto, é o de prever o valor com base nas colunas ‘id’ e ‘idcategoria’.

O Mean Squared Error (Erro Quadrático Médio) de aproximadamente 103818.88 indica que o modelo pode estar enfrentando dificuldades para ajustar-se bem aos dados de teste. Isso significa que as previsões do modelo estão, em média, afastadas dos valores reais por uma quantidade significativa, em termos quadráticos.

As previsões [801.66, 725.04] para os novos dados indicam os valores previstos pelo modelo para os pontos de dados [1, 10] e [2, 15]. Os valores demonstrados pelo modelo são justificáveis pois foram utilizados os dados da API do parceiro, dado que, são dados criados aleatoriamente e podem não fazer sentido. Entretanto, é fundamental ressaltar que o modelo está estruturado para a utilização em data frames reais.

8.3 Métodos de avaliação e validação com o CRISP-DM

O CRISP-DM (Cross Industry Standard Process for Data Mining) é um modelo padrão que define um processo estruturado e organizado para projetos de análise de dados.

O modelo é composto por seis fases:

- Compreensão do Negócio: Nesta fase, é feita uma definição clara do objetivo do projeto, dos recursos necessários e da estratégia para atingir os resultados esperados.
- Compreensão dos Dados: Os dados são coletados, limpos e organizados de forma apropriada.
- Preparação dos Dados: Nesta fase, os dados são preparados para a modelagem.
- Modelagem: São selecionados os modelos que serão utilizados na análise de dados.
- Avaliação: Os resultados são avaliados e refinados até que o modelo alcance o desempenho desejado.
- Implantação: A solução final é implementada e monitorada para garantir que está funcionando de acordo com o esperado.

O CRISP-DM é um modelo iterativo, ou seja, as etapas são realizadas em ciclos até que os resultados esperados sejam alcançados. Isso significa que, após a fase de implantação, a solução deve ser monitorada continuamente e refinada sempre que necessário.

8.4 Proposta de otimização do modelo

Ao analisar os resultados do modelo inicial, identificamos oportunidades para otimização e melhoria do desempenho.

1. Ajuste de Hiperparâmetros:

- Objetivo: O ajuste de hiperparâmetros envolve a modificação dos parâmetros do modelo que não são aprendidos durante o treinamento, mas influenciam o desempenho do modelo.
- Estratégia: Testar diferentes valores para hiperparâmetros, como o número de estimadores na floresta (`n_estimators`), profundidade máxima das árvores (`max_depth`), entre outros. Isso pode ser feito usando técnicas como busca em grade (`Grid Search`) ou busca aleatória (`Random Search`).

2. Engenharia de Recursos (Feature Engineering):

- Objetivo: Identificar e criar novas características que possam melhorar a capacidade do modelo de capturar padrões nos dados.
- Estratégia: Avalia se há outras características no conjunto de dados ou se é possível criar novas características que possam ser mais informativas para a previsão da variável alvo ('value').

3. Outros Algoritmos:

- Objetivo: Explorar diferentes algoritmos de aprendizado de máquina para determinar se algum deles se ajusta melhor ao seu conjunto de dados.
- Estratégia: Experimentar outros modelos, como regressão linear, support vector machines, gradient boosting,

4. Validação Cruzada:

- Objetivo: Avaliar o desempenho do modelo de uma maneira mais robusta, garantindo que ele se generalize bem para dados não vistos.
- Estratégia: Utilizar técnicas de validação cruzada, como k-fold cross-validation, para avaliar a performance do modelo em diferentes conjuntos de treino/teste.

Além disso, é essencial uma análise detalhada dos resultados de cada ajuste, assegurando uma iteração constante e refinada do modelo. Esta abordagem iterativa não apenas visa alcançar a máxima eficácia na previsão da variável alvo, mas também proporciona uma otimização progressiva, o que resulta em um modelo mais robusto e preciso ao longo do tempo.

9. Infográficos

A criação de infográficos é uma prática do design e visualização de dados, que proporciona de uma maneira eficaz a transmissão de informações de forma clara, concisa e atraente. Nesta etapa, os infográficos foram desenvolvidos com o propósito de medir o potencial de consumo por canal, região e categoria. O foco principal é fornecer insights valiosos para a Integration, uma empresa de consultoria, para que possa tomar decisões mais informadas e estratégicas.

9.1 Objetivos

O desenvolvimento destes infográficos tem como objetivos principais fornecer uma análise detalhada e multifacetada sobre o potencial de consumo, canalizando informações específicas e insights açãoáveis em várias dimensões. Primeiramente, buscamos assegurar uma Clarezza de Objetivo nítida: é fundamental que os infográficos sejam imediatamente compreensíveis, destacando sem ambiguidades as informações críticas acerca do potencial de consumo, englobando categorias, canais e regiões específicas.

Em seguida, nosso foco se direciona para a Análise de Canal, Região e Venda, onde o intuito é proporcionar uma visão holística do desempenho dos canais de venda em diferentes regiões. Esta análise é vital para identificar tanto oportunidades quanto desafios existentes, permitindo uma melhor adaptação e estratégia de mercado. Paralelamente, colocamos ênfase no Estado por Consumo Alimentício, oferecendo uma análise especializada sobre os padrões de consumo alimentar. Este aspecto visa destacar as variações significativas de consumo entre os estados, oferecendo uma perspectiva mais granular e específica.

Além disso, é de grande importância a apresentação de dados sobre a Renda Total, que se relaciona diretamente ao potencial de consumo. Esta visão permite uma compreensão mais aprofundada da capacidade financeira das regiões, sendo um

índic平 chave para a tomada de decisão estratégica. Ademais, planejamos incorporar um Perfil do Consumidor por Estado, utilizando um gráfico de teia (spider web) para uma representação visual e intuitiva do perfil do consumidor em diferentes estados, levando em conta uma variedade de atributos relevantes.

Por fim, reconhecemos a importância de uma Avaliação Contínua e aprimoramento dos infográficos. Entendemos que estes são recursos dinâmicos, que devem evoluir constantemente para refletir mudanças e tendências de mercado. Essa avaliação contínua tem como objetivo identificar tanto os pontos fortes quanto os fracos dos infográficos atuais, assegurando sua eficácia e valor contínuo para a Integration. Este processo de aperfeiçoamento contínuo é crucial para manter os infográficos como ferramentas valiosas e relevantes para a análise de mercado.

9.2 Gráficos iniciais com o Power BI

Através de um conjunto diversificado de gráficos disponíveis no [PDF](#), aprimoramos significativamente a compreensão das bases de dados públicos escolhidas (CNPJs, POF e CEP) e da API Cliente - Sale. Esses gráficos facilitam o entendimento de cada base, destacando informações que podem ser decisivas ou irrelevantes na construção dos infográficos finais.

Primeiramente, na Base POF, temos gráficos variados, como o *Gráfico de Barras de Renda Total por Unidade de Federação*, que visa demonstrar a distribuição da renda em diferentes estados. Outro exemplo é a *Tabela com o Local de Aquisição*, mostrando a quantidade de compras em diferentes locais. O *Filtro por Estado* permite uma visualização personalizada, enquanto o *Cartão com o Total de Renda Bruta* oferece uma visão geral da renda, com a opção de filtrar por região.

Além disso, há gráficos que analisam a forma de aquisição, a situação do domicílio entre rural e urbano, e até a distribuição por cor ou raça. A idade e o ano de nascimento dos indivíduos são exibidos em gráficos de linhas, enquanto os padrões de consumo durante a semana e as ocasiões de consumo são detalhados em gráficos de barras.

Já na Base de Dados API Cliente - Sale, os gráficos incluem *Gráficos de Barras* com a contagem de produtos mais vendidos por categoria, permitindo identificar os itens mais populares. Um *Gráfico de Linhas de Série Temporal* ilustra padrões de venda de produtos

ao longo do tempo, e um *Cartão com o Total de Vendas* destaca o volume de vendas. Um Filtro por *Categoria* é providenciado para análises mais específicas.

Cada um desses gráficos desempenha um papel crucial na análise, ajudando a identificar tendências, padrões e insights. Eles são ferramentas essenciais para a compreensão profunda das bases de dados, auxiliando na determinação de quais informações são relevantes para a construção dos infográficos finais, assegurando que as decisões baseadas nesses dados sejam tão informadas e precisas quanto possível.

9.3 Gráficos gerados no Metabase

A análise de dados é uma prática que visa a compreensão de padrões, tendências e oportunidades em diferentes conjuntos de dados. Neste contexto, após as análises iniciais com os infográficos no Power Bi, novos gráficos foram gerados no ambiente Metabase, utilizando correlações de dados e SQL, em que foi possível criar visualizações intuitivas e acessíveis, facilitando a interpretação e a extração de insights. Este conjunto diversificado de infográficos abrange dados provenientes de bases públicas e dados da API do cliente.

9.4 Infográficos Criados no Metabase

Os gráficos gerados no Metabase foram baseados nas VIEWS criadas no Redshift, a partir dos modelos Ensemble. Abaixo estão descritos seus objetivos:

1) O gráfico "Consumo Total por Região" é uma ferramenta poderosa para analisar o volume de consumo em diferentes áreas geográficas. Ele permite às empresas identificar mercados com alta demanda, otimizando estratégias de distribuição e marketing.

2) O "Dieta mais frequente por Estado" oferece uma visão clara das tendências alimentares regionais, ajudando a moldar ofertas de produtos conforme os hábitos locais de consumo, o que é crucial para o planejamento de estratégias de mercado segmentadas.

3) O gráfico "Média de Consumo por Dia da Semana e Tipo de Despesa" revela os padrões de consumo em diferentes dias da semana, divididos por categorias de despesa. Este insight é vital para planejar promoções e ajustar operações para atender às variações na demanda.

4) O "Média de Despesa por Horário de Consumo" fornece uma análise detalhada dos gastos dos consumidores em diferentes momentos do dia, permitindo a otimização de horários de funcionamento e promoções.

5) "Média de gasto por Categoria" e "Média do Valor de Compra por Estado" destacam as preferências de consumo e o poder de compra em várias regiões, oferecendo informações valiosas para estratégias de precificação e distribuição de produtos.

9.5 Relatório de Análise de Eficácia dos Infográficos

Teste de Usabilidade com feedback do usuário: Durante a última sprint retrospective, dia 08/12/2023, o cliente pode ter a primeira visualização do infográfico e pudemos coletar vários feedbacks positivos e negativos. O cliente achou a navegação intuitiva e clara, mas apontou áreas de confusão em gráficos onde as barras estavam supostamente todas com mesmo tamanho, legendas de colunas e etc; Em 'Sugestões de Melhoria' será destacado onde é necessário e interessante de ocorrer mudanças na entrega do infográfico. Em geral, as avaliações revelaram uma satisfação, mas será preciso personalizar e processar (limpar) alguns dados para uma melhor navegação dentro da ferramenta.

9.5.1.Sugestões de Melhoria:

Pensando sobre o visual da infográfico foi apontado a necessidade de facilitar a visualização dos dados nos gráficos. A seguir é apresentado o que é/seria interessante ser alterado em cada gráfico.

Gráfico 1 - Mapa de maior venda em cada estado

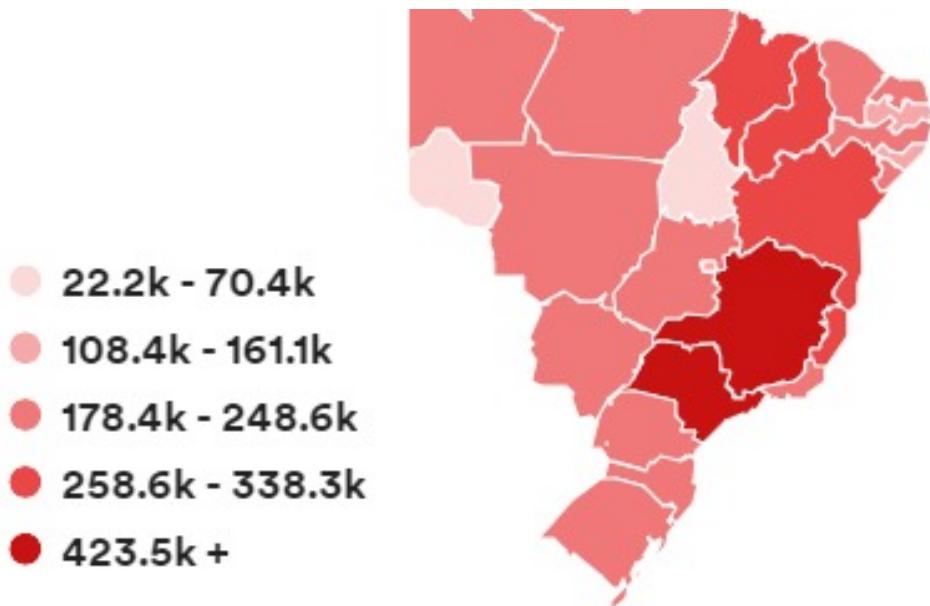


Imagen 36 - Gráfico 1

Ponto de melhoria: Adicionar um título à legenda das cores pois gerou confusão ao identificar o que estava sendo exposto pelas cores. Por exemplo, indicar que quanto mais escuro o tom de vermelho, maior é o número de vendas gerais pelo estado.

Gráfico 2 - Venda de carne por tempo

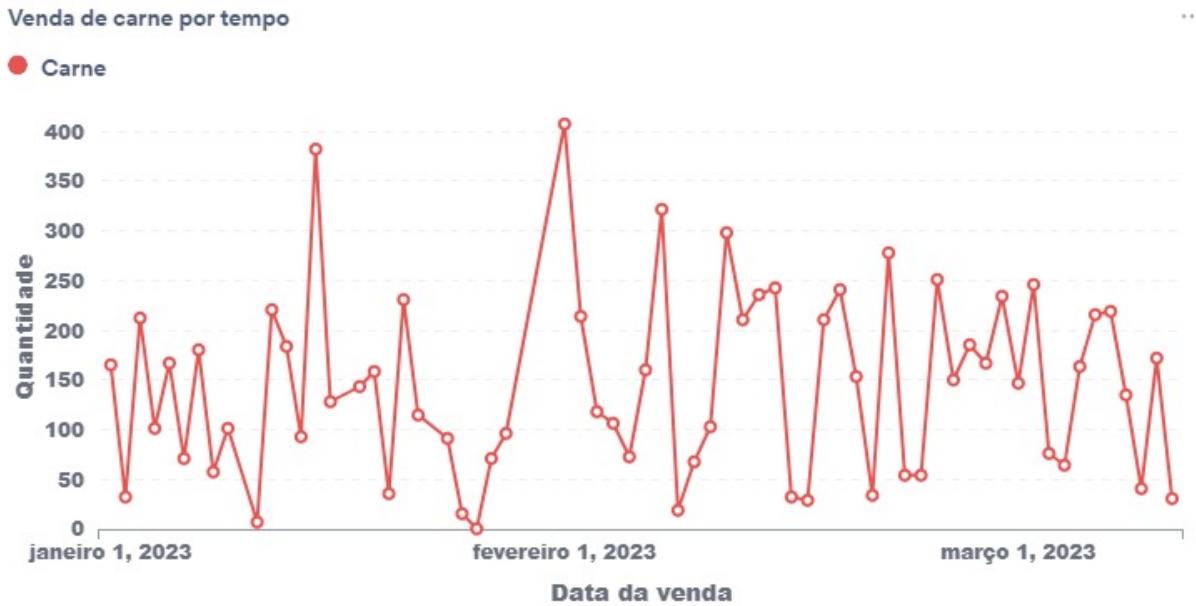


Imagem 37 - Gráfico 2

Ponto de melhoria: Adicionar uma linha que indica a média do número de vendas para facilitar a visualização de outliers.

Gráfico 3 - Média de despesa por horário de consumo

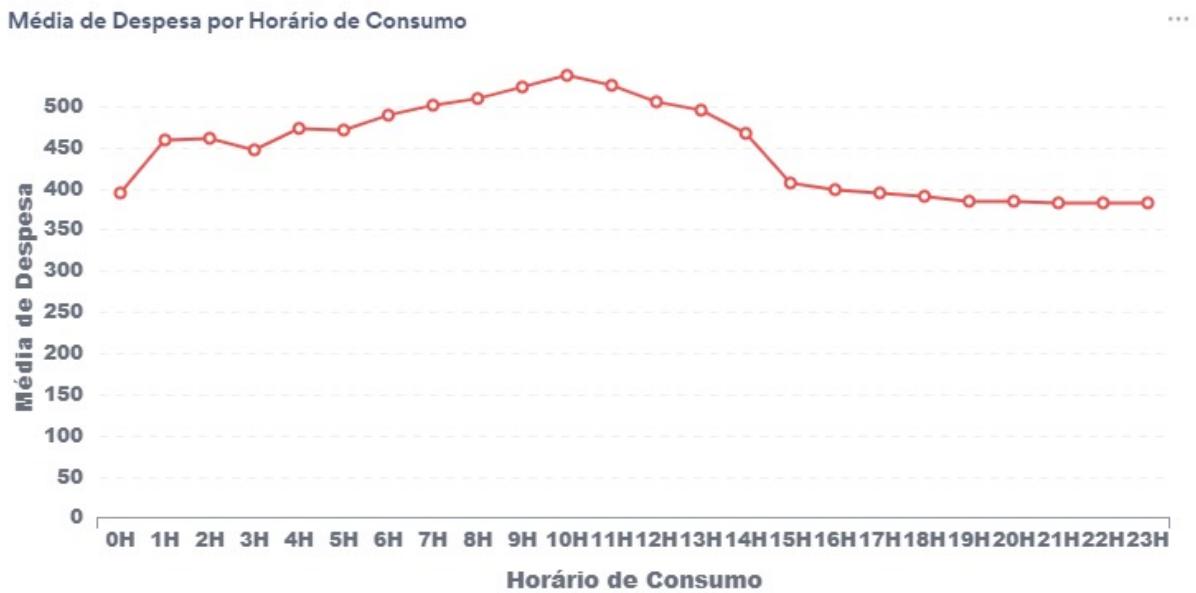


Imagem 38- Gráfico 3

Ponto de melhoria: Adicionar uma linha que indica a média de despesa e facilitar a visualização de outliers.

Gráfico 4 - Refeições mais realizadas

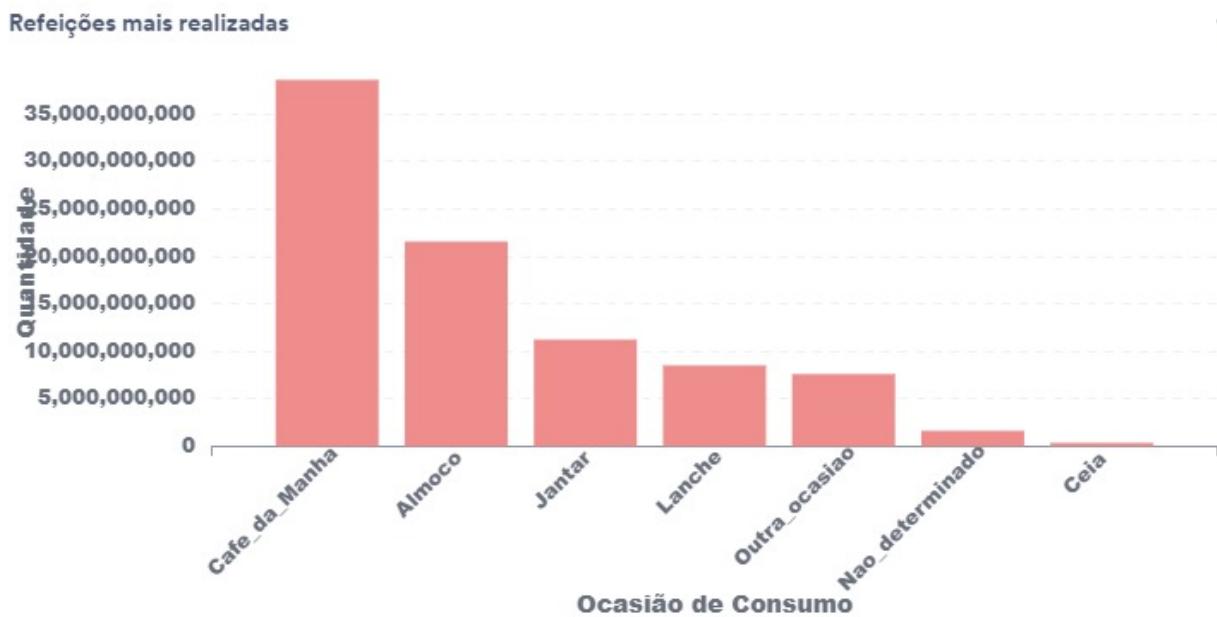


Imagen 39 - Gráfico 4

Ponto de melhoria: Melhorar a escrita da legenda da coluna Y (número de refeições).

Como são números grandes ficam difíceis de visualizar e acabam poluindo o design.

Gráfico 5 - Categoria mais vendida

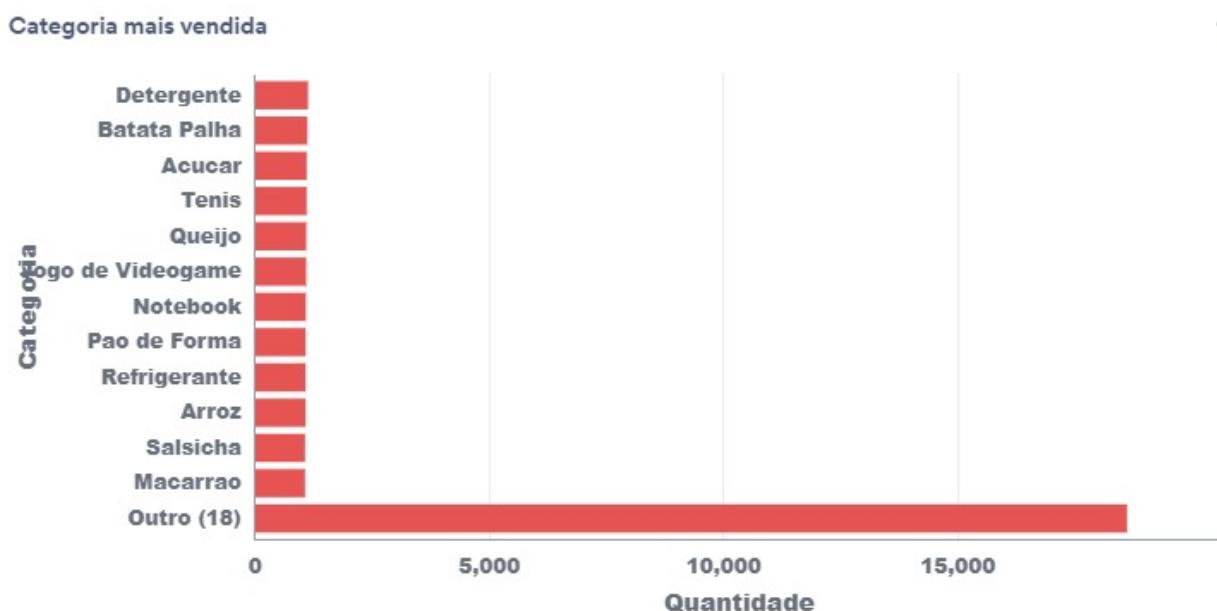


Imagen 40 - Gráfico 5

Ponto de melhoria: Facilitar a visualização da diferença de tamanho entre as barras. Pode ser feito retirando talvez a coluna 'outros'.

Gráfico 6 - Quantidade consumida por hora

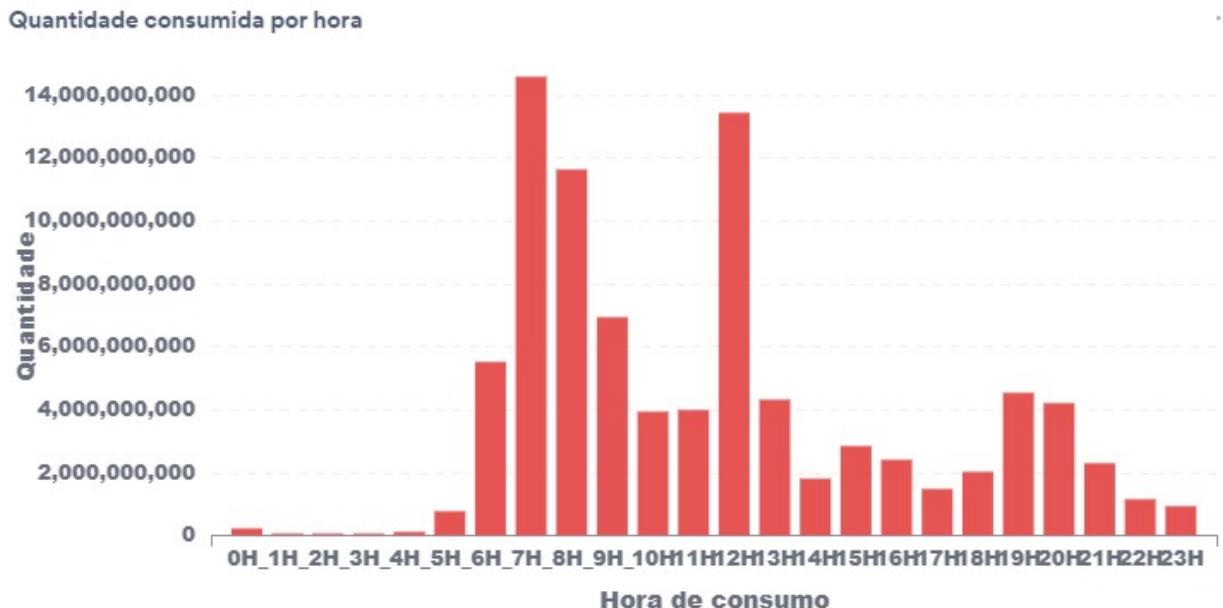


Imagen 41 - Gráfico 6

Ponto de melhoria: Adicionar uma linha que indica a média de quantidade consumida para facilitar a visualização de outliers.

Gráfico 7 - Consumo por local de refeição em dias atípicos

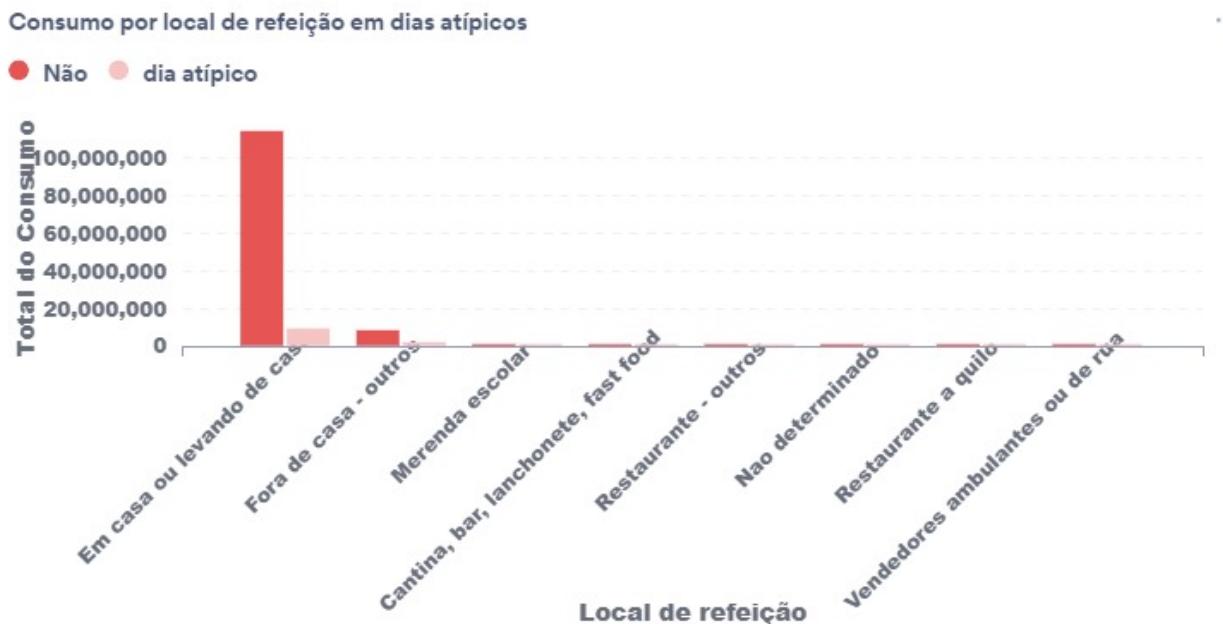


Imagen 42 - Gráfico 7

Ponto de melhoria: Melhorar a escrita da legenda da coluna Y (Total de consumo). Como são números grandes ficam difíceis de visualizar e acabam poluindo o design. Além disso, estudar se é necessário manter as colunas que estão muito pequenas como 'restaurante a quilo' ou 'não identificado', talvez deixar só 3 colunas principais.

Gráfico 8 - Média de valor de compra por estado

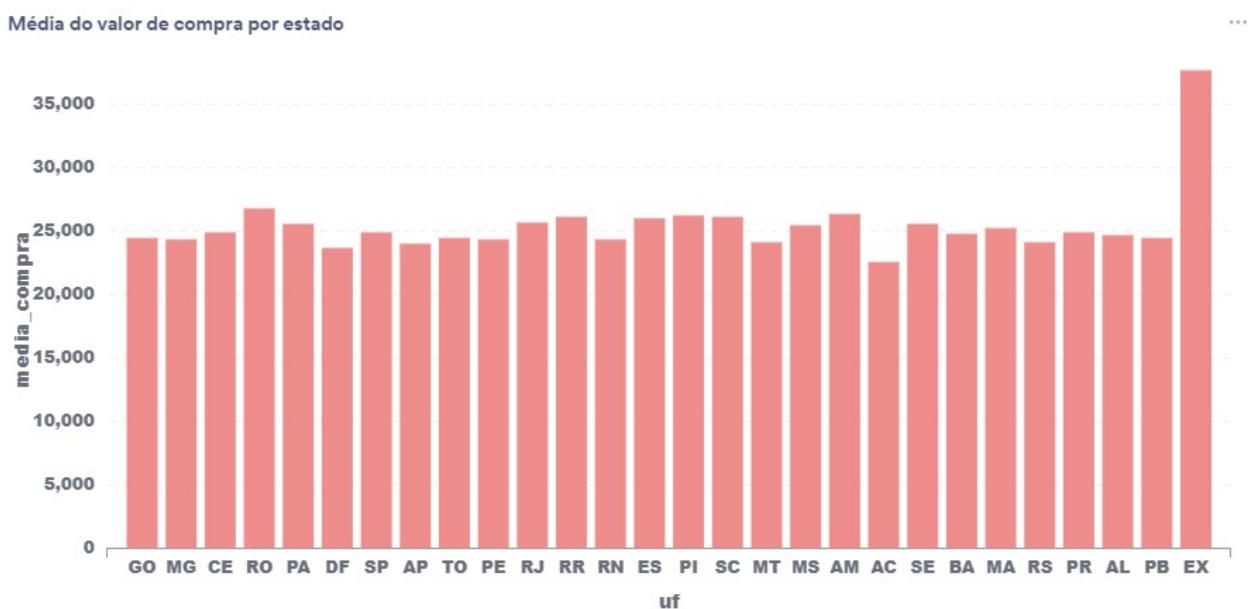


Imagen 43 - Gráfico 8

Ponto de melhoria: Retirar a coluna 'EX' pois não faz sentido para análises de território brasileiro. Além disso é necessário destacar de alguma forma quais seguem a média e quais se sobressaem, pode-se fazer isso com uma linha talvez. Por outro lado, pode se imaginar de adicionar mais um filtro dentro deste mesmo gráfico, criando clusters de padrões por estado, por exemplo colunas que estiverem da mesma cor consomem mais carne e colunas de outra cor consomem mais notebooks.

Gráfico 9 - Tipo de dieta por estado

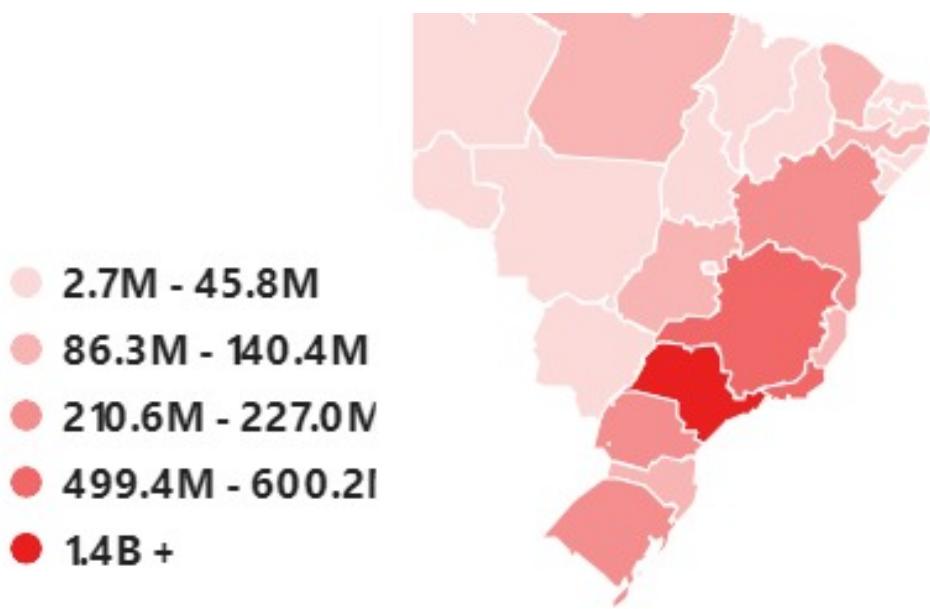


Imagen 44 - Gráfico 9

Ponto de melhoria: Adicionar um título à legenda das cores pois gerou confusão ao identificar o que estava sendo exposto pelas cores. Por exemplo, indicar que quanto mais escuro o tom de vermelho, maior é o número de consumo geral pelo estado.

9.5.2.Conclusão

Em conclusão, o cliente gostou da entrega e achou que temos que aprimorar no tratamento e limpeza dos dados durante a criação dos gráficos até a próxima e última entrega. As sugestões propostas visam otimizar a experiência do usuário, garantindo que o infográfico cumpra de maneira efetiva seu propósito de facilitar a criação de estratégias Go-to-market. A implementação dessas melhorias promoverá um impacto significativo, proporcionando aos usuários uma ferramenta mais eficaz e personalizada para a análise de informações complexas e massivas.

10. Análise financeira

10.1. Receita

A proposta elaborada não possui uma fonte de receita direta. Para realizar a avaliação financeira detalhada, seria necessário obter acesso a informações altamente confidenciais da Integration, as quais não nos foram disponibilizadas.

10.2. Orçamento total

Cada um dos membros que compõem o grupo dispõe de um saldo individual de 100 dólares em créditos na plataforma da AWS, somando assim um montante coletivo de 600 dólares destinado para utilização dentro da plataforma AWS. O orçamento foi disponibilizado para que a equipe consiga fazer testes reais dentro da plataforma.

Contudo, foi preciso adicionar novas contas com créditos, uma vez que o total coletivo de 600 dólares não se mostrou adequado para cobrir os testes ao longo das sprints.

10.3. Custos

10.3.1. Métodos de pagamento AWS

Existem 4 métodos de pagamento disponíveis na AWS (Compute Savings Plans, On demand, Spot Instances e EC2 Instance Savings Plans). No caso do projeto elaborado, por se tratar de um MVP, a melhor opção é o pagamento sob demanda porque ele oferece a flexibilidade necessária para ajustar os recursos de acordo com as necessidades variáveis do projeto em sua fase inicial. O pagamento sob demanda permite que você comece com uma base mínima de recursos e aumente gradualmente à medida que a demanda do MVP cresce.

O modelo de pagamento sob demanda é especialmente vantajoso quando se trabalha em um MVP, onde as estimativas iniciais de uso podem variar e evoluir rapidamente à medida que o projeto é refinado e desenvolvido. Não há necessidade de se comprometer com um contrato de longo prazo ou investir em recursos que possam não ser completamente utilizados no estágio inicial. Além disso, a abordagem de pagamento sob demanda alinha-se perfeitamente com os princípios ágeis e iterativos comuns na construção de um MVP.

É importante monitorar de perto seus custos e considerar a transição para outras opções de pagamento, como Instâncias Reservadas ou Planos de Economia de Computação no caso de uso do projeto por parte da empresa. Essas opções podem proporcionar economias substanciais a longo prazo, à medida que você adquire mais previsibilidade em relação ao uso de recursos.

10.3.2.Custo dos Serviços utilizados

Primeiramente, é essencial realizar a enumeração dos serviços da AWS que foram empregados no projeto. São eles:

AWS Lambda:

- O AWS Lambda é um serviço serverless que permite a execução de código de maneira escalonável sem a necessidade de gerenciar a infraestrutura subjacente. A cobrança ocorre com base no número de execuções e no tempo de processamento do código.

AWS Redshift:

- O Amazon Redshift é um serviço de data warehousing na nuvem, projetado para executar análises em grandes conjuntos de dados. A cobrança está associada ao tamanho do cluster e à quantidade de dados processados em consultas.

Apache Spark:

- Embora o Apache Spark não seja exclusivo da AWS, é uma estrutura de processamento de dados distribuída frequentemente utilizada na nuvem para análise e processamento de dados em larga escala. Os custos podem estar relacionados ao uso de recursos computacionais e de armazenamento.

AWS Quicksight:

- O Amazon QuickSight é um serviço de visualização de dados que permite criar dashboards interativos. A cobrança é baseada na quantidade de usuários ativos e no volume de dados processados para visualização.

AWS CloudWatch:

- O Amazon CloudWatch é um serviço de monitoramento e observabilidade que coleta e rastreia dados operacionais. Os custos estão associados à

quantidade de métricas, logs e eventos monitorados, bem como ao armazenamento desses dados.

AWS KMS (Key Management Service):

- O AWS Key Management Service (KMS) é um serviço de gerenciamento de chaves de criptografia. A cobrança está relacionada ao número de operações de criptografia, criação e gerenciamento de chaves.

VPC (Virtual Private Cloud):

- A AWS não cobra pela criação e configuração da Amazon Virtual Private Cloud (VPC) em si. No entanto, a utilização de recursos dentro da VPC, como instâncias EC2, bancos de dados e outros serviços, está sujeita às taxas normais associadas a esses recursos. A cobrança da VPC é indireta e depende do consumo de recursos alocados dentro dela.

10.3.3. Mapeamento de custos 12 meses AWS:

Serviço	Detalhes Técnicos	Custo Mensal (Dólares)	Custo para 12 Meses (Dólares)	Resumo da Configuração
AWS Lambda	Execução de código serverless, escalonável	\$0.00	\$0.00	Arquitetura (x86), Modo de invocação (Em buffer), Quantidade de armazenamento temporário alocada (512 MB)

AWS Redshift	Data warehousing na nuvem para análise	\$1,263.63	\$15,163.56	Nós (1), Tipo de instância (ra3.xlplus), Utilização (somente sob demanda) (100 %Utilized/Month), Modelo de preço (OnDemand)
AWS Quicksight	Serviço de visualização de dados	\$31.40	\$376.80	Número de dias úteis por mês (22), Capacidade SPICE em gigabytes (GB) (10), Número de autores (1), Número de leitores (6)
Bucket S3	Serviço de monitoramento e observabilidade	\$17,28	\$207,36	
AWS KMS	Serviço de gerenciamento de chaves de criptografia	\$12.00	\$144.00	Número de chaves mestra de cliente (CMK) gerenciadas pelo cliente (6), Número de solicitações simétricas (2,000,000)

VPC	Rede virtual isolada na AWS	\$0.00	\$0.00	Dias úteis por mês (22)
-----	-----------------------------	--------	--------	-------------------------

Valor Total para 12 Meses = \$15,891.72

Valor Total Mensal: \$ 1,292.91

Cotação do dólar: R\$ 4,90 (06/12/2023)

10.3.4. Análise comparativa entre Amazon Web Services (AWS) e Microsoft Azure

Mapeamento de custos 12 meses Azure:

Serviço	Detalhes Técnicos	Custo Mensal (Dólares)	Custo para 12 Meses (Dólares)	Resumo da Configuração
Azure Functions	Serverless, escalável	\$0.00	\$0.00	Aplicações e funções executadas sob demanda na plataforma Azure
Azure Synapse Analytics	Data warehousing na nuvem para análise	\$507.47	\$6,089.64	Armazenamento e análise de grandes conjuntos de dados na Azure
Storage Accounts	Armazenamento em nuvem	\$7.23	\$86.76	Armazenamento escalável e seguro na nuvem da Microsoft Azure

API Management	Gerenciamento de APIs	\$3.23	\$38.76	Plataforma para criação, publicação e gerenciamento de APIs
Azure Virtual Machines	Máquinas Virtuais na Azure	\$12.61	\$151.32	Implementação e gerenciamento de máquinas virtuais na nuvem

Valor Total para 12 Meses = \$6,366,36

Valor por mensal = \$530,53

Cotação do dólar: R\$ 4,90 (06/12/2023)

10.3.5.Custos de desenvolvimento

Também é fundamental calcular os custos associados à prestação do serviço de desenvolvimento da aplicação a fim de garantir uma estimativa abrangente.

1. Duração do projeto:

O tempo de desenvolvimento da solução foi de 10 semanas (divididas em 5 Sprints).

2. Salário de cada desenvolvedor:

Baseado na média salarial de um gerente de projetos júnior, conforme informações disponíveis em https://www.glassdoor.com.br/Sal%C3%A1rios/gerente-de-projetos-j%C3%BAnior-sal%C3%A1rio-SRCH_KO0,26.htm

3. Custos relacionados à manutenção do projeto:

Neste caso, o custo para manutenção é composto pelos custos cobrados pela AWS sobre a alocação da infraestrutura (armazenamento, poder computacional, etc.) em nuvem utilizada no projeto.

4. Horas totais:

Considerando o plano da faculdade para este módulo e descontando os dias de encontros com o cliente e as Sprints Planning, temos 232h 30min disponíveis para a realização do estudo e desenvolvimento do projeto.

5. Remuneração por hora:

Com base no salário total de cada desenvolvedor e na quantidade total de horas trabalhadas, a remuneração por hora para cada gerente de projetos júnior é de R\$63,66.

6. Custo total de desenvolvimento do projeto:

A soma de todos os custos com salários resulta em um custo total de R\$91.142,10.

Item	Descrição	Valor
Duração do projeto	10 semanas	-
Número de integrantes no time	6 desenvolvedores	-
Salário mensal de cada desenvolvedor (júnior)	Média salarial do mercado	R\$ 7.400,00
Horas disponíveis por desenvolvedor		232h 30min
Remuneração por hora por desenvolvedor	Salário mensal 2 meses / horas totais	R\$ 63,66
Custo total de desenvolvimento do projeto	(Remuneração por hora Horas totais) Número de desenvolvedores	R\$ 91.142,10

11. Impacto Ético

11.1. Introdução

Conforme avançamos a era da informação, somos confrontados com a grande capacidade de coletar, processar e analisar extensos conjuntos de dados, que é conhecido como Big Data. Este capítulo propõe mostrar os impactos do Big Data na sociedade e no meio ambiente, inicialmente concentrando-se na esfera crítica da privacidade e proteção de dados, alinhada às exigências normativas, como a Lei Geral de Proteção de Dados (LGPD).

11.2. Privacidade e Proteção de Dados

11.2.1 Coleta de Informações Pessoais

A coleta massiva de dados é uma característica central do Big Data. Organizações agora têm a capacidade de extrair informações pessoais significativas de fontes variadas, como redes sociais, transações online, dispositivos conectados, entre outros. Essa prática levanta questões substanciais sobre a privacidade dos indivíduos [Smith, 2019].

11.2.2 Armazenamento e Processamento de Dados

O armazenamento eficiente e o processamento rápido de grandes volumes de dados são fundamentais para as aplicações de Big Data. Porém, a centralização dessas informações em grandes repositórios de dados pode ser alvo de preocupações relacionadas à segurança e privacidade. Vulnerabilidades nesses sistemas podem resultar em acessos não autorizados, expondo dados sensíveis [Jones et al., 2020].

11.2.3 Uso de Informações Pessoais

O uso de dados pessoais para análises preditivas e personalizações levanta questões éticas e morais. Empresas podem segmentar consumidores com base em seu

comportamento registrado, o que pode influenciar decisões de consumo e moldar a experiência do usuário de maneiras sutis. O desafio reside em garantir que o uso dessas informações seja transparente e ético [Garcia, 2018].

11.2.4 Conformidade com a LGPD e Regulamentações de Privacidade

A Lei Geral de Proteção de Dados (LGPD) no Brasil estabelece diretrizes claras para a coleta, armazenamento e processamento de dados pessoais. O Big Data, ao lidar com grandes volumes de informações, deve aderir estritamente a essas regulamentações para proteger a privacidade dos cidadãos. Falhas na conformidade podem resultar em penalidades significativas [Brasil, 2018].

11.3. Equidade e justiça

11.3.1 Diversidade nos Conjuntos de Dados

Promover a diversidade nos conjuntos de dados utilizados no treinamento de algoritmos é crucial para mitigar disparidades. A inclusão representativa de dados provenientes de diferentes grupos étnicos, sociais e demográficos contribui para a criação de modelos mais equitativos [Crawford, 2017].

11.3.2 Técnicas de Correção de Viés

A aplicação de técnicas de correção de viés, como reamostragem ponderada e modificação de algoritmos, é fundamental para garantir que os modelos gerados não perpetuem injustiças. Estas técnicas visam ajustar os resultados dos algoritmos para refletir de maneira mais precisa a realidade [Hardt et al., 2016].

11.3.3 Transparência e Responsabilidade

Promover a transparência nos algoritmos é essencial para entender como as decisões são tomadas e identificar possíveis fontes de viés. Além disso, garantir que os desenvolvedores e usuários sejam responsáveis pela implementação e uso ético de algoritmos contribui para uma abordagem mais justa em relação aos dados [Diakopoulos, 2016].

11.4. Transparência em Projetos de Big Data

11.4.1 Acesso Claro às Informações Relevantes

A transparência em projetos de Big Data implica que todas as partes interessadas tenham acesso claro e compreensível às informações relevantes. Isso inclui detalhes sobre quais dados estão sendo coletados, como serão usados, quem terá acesso a eles e quais são as potenciais ramificações para os indivíduos envolvidos [Mittelstadt et al., 2016].

11.4.2 Garantindo Conformidade com Normas e Regulamentações

A transparência também está diretamente ligada à conformidade com normas e regulamentações, como a Lei Geral de Proteção de Dados (LGPD). Documentar e comunicar claramente como os dados serão tratados, garantindo que as práticas estejam em conformidade com padrões éticos e legais, é fundamental [Dignum et al., 2018].

11.4.3 Obtendo Consentimento Adequado

O consentimento informado em projetos de Big Data exige que os indivíduos estejam cientes e concordam explicitamente com a coleta e uso de seus dados. Este processo deve ser transparente, sem ambiguidades, e oferecer aos participantes a oportunidade de fazer escolhas informadas [Mittelstadt et al., 2016].

11.4.4 Consentimento Contínuo

Dada a natureza dinâmica do Big Data, onde os objetivos e métodos de análise podem evoluir, é vital adotar uma abordagem de consentimento contínuo. Isso implica informar os participantes sobre quaisquer mudanças significativas no processamento de dados e permitir que eles reavaliem e atualizem seu consentimento conforme necessário [Dignum et al., 2018].

11.4.5 Governança de Dados

A governança de dados, discutida por Dignum et al. (2018), oferece uma estrutura para garantir a transparência e o consentimento informado. A governança abrange políticas, processos e tecnologias que ajudam a garantir que a coleta e o uso de dados ocorram de maneira ética e alinhada com as expectativas dos participantes.

11.5. Estratégias para Garantir Transparência e Consentimento em Big Data

11.5.1 Comunicação Clara e Acessível

Adotar estratégias de comunicação clara e acessível é crucial para garantir que as informações relevantes cheguem a todas as partes interessadas. Isso inclui o uso de linguagem simples e a oferta de canais de comunicação eficazes para esclarecer dúvidas [Mittelstadt et al., 2016].

11.5.2 Plataformas Interativas de Consentimento

A implementação de plataformas interativas de consentimento, que permitem aos participantes navegar pelas opções de privacidade de maneira personalizada, pode facilitar o processo de consentimento informado [Dignum et al., 2018]. Essas plataformas podem ser projetadas para garantir que os participantes compreendam totalmente as implicações de suas escolhas.

11.6.Responsabilidade social

A responsabilidade social em projetos de Big Data é crucial para avaliar e mitigar os potenciais impactos sociais, positivos e negativos, sobre comunidades e o meio ambiente. Este capítulo explora como a consideração responsável desses impactos pode contribuir para a sustentabilidade e alinhamento com Objetivos de Desenvolvimento Sustentável (ODS), ao mesmo tempo em que destaca a importância de tratar dados sensíveis, como telefones e e-mails, com responsabilidade ética.

11.6.1 Objetivos de Desenvolvimento Sustentável (ODS)

Os Objetivos de Desenvolvimento Sustentável, propostos pela ONU em 2015, fornecem uma estrutura global para orientar práticas que promovem a sustentabilidade e o bem-estar social. A avaliação do alinhamento de projetos de Big Data com esses objetivos é crucial para a responsabilidade social [United Nations, 2015].

11.6.2 Avaliação de Impacto Social

Incorporar uma avaliação de impacto social no início do ciclo de vida do projeto é essencial. Isso envolve a identificação proativa de possíveis efeitos sobre comunidades e a implementação de estratégias para maximizar impactos positivos e mitigar impactos negativos [Floridi et al., 2018].

11.6.3 Educação e Conscientização

Promover a educação e conscientização sobre responsabilidade social entre as equipes de projeto e partes interessadas é fundamental. Isso inclui treinamento sobre a importância de tratar dados sensíveis com ética e considerar os impactos sociais em todas as fases do projeto [Mittelstadt et al., 2016].

11.7. Tratamento Ético de Dados Sensíveis

11.7.1 Dados Pessoais: Telefones e E-mails

O tratamento ético de dados sensíveis, como telefones e e-mails, é uma peça fundamental da responsabilidade social. É imperativo adotar medidas rigorosas de segurança e privacidade para garantir que essas informações não sejam exploradas de maneira inadequada, respeitando regulamentações como a LGPD [Brasil, 2018].

11.7.2 Evitar Violações de Privacidade

A responsabilidade social inclui o compromisso de evitar violações de privacidade relacionadas a dados pessoais sensíveis. Isso implica a implementação de práticas de segurança robustas, criptografia e o uso ético dessas informações apenas para os fins explicitamente comunicados e consentidos pelos indivíduos [Mittelstadt et al., 2016].

11.8. Viés Algorítmico e Discriminação em Projetos de Big Data

11.8.1 Riscos de Viés Algorítmico

O viés algorítmico, muitas vezes derivado de conjuntos de dados desbalanceados ou representações inadequadas, pode resultar em decisões discriminatórias automatizadas. Isso afeta negativamente certos grupos, exacerbando desigualdades sociais preexistentes [Barocas & Hardt, 2019].

11.8.2 Discriminação e Exclusão Involuntária

Algoritmos de Big Data podem inadvertidamente contribuir para a discriminação, excluindo involuntariamente certos grupos. Isso pode ocorrer devido a preconceitos

existentes nos dados utilizados para treinamento ou a lacunas na representação de certas comunidades [O'Neil, 2016].

11.9. Estratégias para Mitigar Viés e Discriminação

11.9.1 Auditoria de Algoritmos

A auditoria de algoritmos, um processo crítico proposto por Barocas e Hardt (2019), envolve a análise sistemática dos resultados do algoritmo para identificar e corrigir possíveis viés. Esta prática permite uma compreensão mais profunda dos padrões de tomada de decisão e sua justiça.

11.9.2 Diversidade nos Conjuntos de Dados

A inclusão de dados diversificados é uma estratégia-chave para mitigar o viés algorítmico. Diversificar os conjuntos de dados utilizados no treinamento ajuda a garantir representatividade, evitando a sub-representação ou marginalização de certos grupos [Diakopoulos, 2016].

11.10. Considerações Finais

O avanço do Big Data apresenta um cenário misto de oportunidades e desafios éticos. A proteção da privacidade se destaca como um ponto crucial, considerando as implicações profundas e variadas desse fenômeno tecnológico. A busca por equidade na aplicação de algoritmos, mitigando viés, e a adoção de práticas socialmente responsáveis ganham destaque na construção de um ecossistema digital ético e sustentável. A transparência nas práticas, a obtenção de consentimento informado e a implementação de sólidas estruturas de governança surgem não apenas como medidas de conformidade, mas como alicerces essenciais para a integração ética e eficaz do Big Data na sociedade atual. A colaboração conjunta entre diversos segmentos da sociedade, incluindo a comunidade acadêmica, legisladores e setor privado, revela-se como um elemento determinante para a estruturação de um futuro onde as inovações propiciadas pelo Big Data sejam realizadas de maneira compatível com princípios éticos, preservando direitos individuais e promovendo o bem-estar coletivo.

12. Plano de Comunicação

12.1 Objetivo

O principal objetivo do plano de comunicação é assegurar alinhamento e compreensão entre todas as partes envolvidas no projeto de BIG Data deste módulo, incluindo a empresa Integration e o time Inteli (Orientador e Professores). A comunicação será focada em destacar a relevância do projeto, seus objetivos e benefícios esperados para a Integration Consulting e seus clientes. Além disso, será enfatizada a entrega final do projeto, que incluirá um infográfico e a entrega do cubo de dados. O plano visa garantir que a comunicação sobre o projeto de Big Data seja consistente, comprehensível e alinhada aos objetivos estratégicos da Integration Consulting.

Relevância do Projeto:

- Explorar a importância estratégica do projeto para a Integration Consulting, ressaltando como a análise de BIG Data pode fornecer insights valiosos para a tomada de decisões estratégicas e operacionais.
- Destacar a capacidade do projeto em posicionar a Integration Consulting como uma líder inovadora no uso de tecnologias emergentes para solucionar desafios de negócios complexos.

Objetivos do Projeto:

- Esclarecer os objetivos específicos do projeto, incluindo a criação de um pipeline de Big Data baseado na AWS, análises estatísticas em um datalake/data warehouse, e a produção de um infográfico informativo.
- Demonstrar como esses objetivos contribuem diretamente para a missão mais ampla da Integration Consulting de oferecer soluções eficazes e diferenciadas aos seus clientes.

Benefícios Esperados:

- Comunicar os benefícios tangíveis que a Integration Consulting e seu cliente distribuidor podem esperar alcançar por meio da implementação bem-sucedida do projeto, como:
 - aprimoramento de estratégias de marketing;
 - eficiência operacional;
 - aumento da compreensão do comportamento do consumidor.
- Destacar que os benefícios não se limitam ao curto prazo, mas têm implicações a longo prazo para a vantagem competitiva do cliente.

Entrega do Projeto:

- Criação de um infográfico, que incluirá visualizações detalhadas de:
 - categorias, canais e regiões com potencial de consumo.
- Esse infográfico será uma ferramenta para comunicar descobertas de maneira acessível e impactante.
- Entrega do cubo de dados, evidenciando como essa estrutura tridimensional permite uma análise multifacetada e aprofundada dos dados, proporcionando insights detalhados para decisões estratégicas.

12.2 Stakeholders

- Equipe de Desenvolvimento do Projeto (Inteli).
- Administração da Integration Consulting.
- Equipe Técnica da Integration Consulting.
- Cliente Distribuidor.
- Orientador e Professores do Inteli.

12.3 Mensagens Chave

Equipe de Desenvolvimento do Projeto (Inteli):

- Construção e monitoramento do projeto.
- Progresso e desafios durante o desenvolvimento.
- Importância da replicabilidade da solução em casos semelhantes.

Alta Administração e Equipe Técnica da Integration Consulting:

- Alinhamento estratégico do projeto com os objetivos da empresa.
- Antecipação dos benefícios esperados.

Cliente Distribuidor:

- Transparência sobre o progresso do projeto.
- Destaque para a entrega final: infográfico e cubo de dados.

Orientador e Professores do Inteli:

- Contribuição para o aprendizado prático dos estudantes.
- Envolvimento na entrega do projeto como parte integrante do desenvolvimento acadêmico.

12.4 Canais de Comunicação:

- **Equipe de Desenvolvimento do Projeto (Inteli):**
 - Realização da Daily em grupo;
 - Planning todo o início de sprint;
 - Reuniões semanais para discussão de progresso e desafios;
 - Retrospective todo final de sprint;
 - Status Report ao final das apresentações para o cliente;
 - Alinhar o projeto de acordo com os feedbacks do cliente a cada apresentação;
 - Apresentar os incrementos a cada sprint quando realizado de acordo com a expectativa do cliente;
 - Plataforma slack para comunicação com o grupo, professores e orientadores.
- **Alta Administração e Equipe Técnica da Integration Consulting:**
 - Apresentações em reuniões mensais.
 - Comunicação com o orientador de sala e professores com atualizações detalhadas.
- **Cliente Distribuidor:**
 - Reuniões para revisão do progresso.
- **Orientador e Professores do Inteli:**
 - Reuniões diárias para atualizações e feedback.
 - Apresentações a cada final de sprint do projeto.

12.5 Plano de Implementação

- **Sprint 1 (Semana 1-2):**
 - Reuniões de kick-off com a equipe de desenvolvimento, cliente e orientador para alinhar expectativas e leitura/entendimento sobre o TAPI.
 - **Desenvolvimento do artefato de UX:** Criação de Persona, História do Usuário, Mapa de Jornada de Usuário.
 - **Desenvolvimento do artefato de Negócio:** Canvas Proposta de Valor, Total Addressable Market, Service Addressable Market, e Service Obtainable Market e Matriz de Risco.
 - **Desenvolvimento do artefato de Programação:** Arquitetura de Ingestão de Dados do Parceiro.
 - Validação dos artefatos produzidos para o projeto com o cliente nas apresentações de cada final de sprint.
- **Sprint 2 (Semana 3-4):**
 - Iniciar o desenvolvimento do projeto com foco na configuração inicial do pipeline de Big Data.

- Comunicação regular com a equipe técnica da Integration Consulting e orientador para garantir o alinhamento do projeto.
- Prototipação em Baixa fidelidade da visualização de dados do projeto, com foco na experiência do usuário.
- Estrutura de Ingestão de dados com Armazenamento.
- Alinhamento e validação da arquitetura e estrutura do projeto de Big Data, assim como também do wireframe, com o cliente na apresentação dessa sprint 2.
- **Sprint 3 (Semana 5-6):**
 - Desenvolvimento de Data Lake/Data Warehouse alimentado pelo processo de ETL (Extração, Transformação e Carga), com o objetivo de criar uma estrutura de armazenamento centralizada que permita a ingestão, transformação e disponibilização eficiente dos dados para análises e tomadas de decisão.
 - Documentação de análise de impacto ético.
 - Validação da nossa estrutura AWS com o cliente e alinhamento para os próximos passos durante a apresentação da sprint 3.
- **Sprint 4 (Semana 7-8):**
 - Criação de Infográfico e relatório de análise de eficácia do mesmo com sugestão de melhorias.
 - Modelo de Ensemble com Processamento em Big Data.
 - Análise financeira do projeto.
 - Plano de comunicação.
 - Validação e alinhamento na apresentação com o cliente, referente à sprint 4, sobre o infográfico que estamos desenvolvendo e os dados que estamos apresentando e fazendo correlações.
- **Sprint 5 (Semana 9-10):**
 - Entrega final do Pipeline de Big Data na AWS.
 - Documentação completa da solução.
 - Apresentação Final.

12.6 Medidas de Sucesso

- Retrospective e status report.
- Métricas de desempenho: Throughput, WIP, lead time, para equilíbrio e comunicação entre os membros da equipe.
- Documentação, Pipefy, Github.
- Nível de compreensão e incremento no projeto a partir dos feedbacks dos stakeholders.
- Conclusão de tarefas do projeto no prazo.
- Satisfação do cliente com insights gerados.

- Desenvolvimento de projetos em boas práticas e padrões.
- Feedbacks e apresentações com o cliente.

12.7 Feedback e Ajustes

- Feedbacks fornecidos após apresentações com o cliente.
- Incremento dos feedbacks recebidos.
- Canal de comunicação contínuo através de plataformas de comunicação, como por exemplo o slack.
- Revisões semanais do plano de comunicação para ajustes conforme necessário.

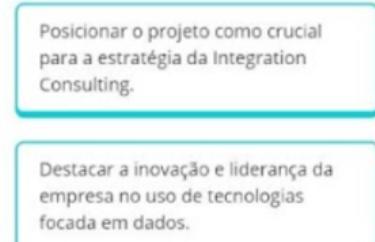
Target audience | 5 ...



Objectives | 3 ...



Positioning | 2 ...



+

+

+



13. Conclusões

A conclusão deste projeto abrange a jornada desde a identificação do problema até a implementação prática de uma solução robusta e adaptável no contexto de Big Data. Ao longo das diversas etapas, desde a compreensão do problema até a análise financeira e considerações éticas, o projeto revelou-se uma abordagem abrangente e integrada para atender às necessidades do parceiro de negócios.

Iniciando com a compreensão aprofundada do problema, passando pela definição clara de objetivos, personas e jornadas do usuário, o projeto forneceu uma base sólida para o desenvolvimento de um sistema de Big Data eficiente. A análise de experiência do usuário, através de personas e user stories, ajudou a direcionar o desenvolvimento, garantindo que a solução estivesse alinhada com as expectativas e necessidades dos usuários finais.

A arquitetura macro detalhada, desde os componentes até a análise exploratória dos dados, destacou a complexidade e a abrangência do sistema proposto. A implementação do Data Warehouse, a curadoria dos dados e a aplicação de modelos ensemble reforçaram a capacidade do pipeline de processar grandes volumes de dados de forma eficaz, proporcionando insights valiosos para a tomada de decisões.

A fase de infográficos trouxe uma dimensão visual às análises, permitindo uma compreensão intuitiva dos dados. O plano de comunicação e o impacto ético enfatizaram a importância de

considerações cruciais, como privacidade, transparência e responsabilidade social, na implementação de projetos de Big Data.

A análise financeira proporcionou uma visão detalhada dos custos associados ao projeto, incluindo uma comparação entre Amazon Web Services (AWS) e Microsoft Azure. Essa seção é crucial para garantir a sustentabilidade financeira do projeto e otimizar os recursos disponíveis.

Logo, podemos afirmar que o pipeline de Big Data desenvolvido não apenas atendeu às necessidades atuais de análise de dados do parceiro de negócios, mas também estabeleceu uma base sólida e adaptável para futuras evoluções.

14. Referências

McKinsey. Transformações digitais no Brasil: insights sobre o nível de maturidade digital das empresas no país. Disponível em: <<https://www.mckinsey.com/br/our-insights/transformacoes-digitais-no-brasil>>. Acesso em: 25/10/2023.

Vivo Meu Negócio. Varejo alimentar: tendências e expectativas. Disponível em: <<https://vivomeunegocio.com.br/bares-e-restaurantes/gerenciar/varejo-alimentar/>>. Acesso em: 25/10/2023.

Statista. Business Intelligence Software - Brazil. Disponível em: <<https://www.statista.com/outlook/tmo/software/enterprise-software/business-intelligence-software/brazil>>. Acesso em: 25/10/2023.

INVESTSP. Setores de Negócios: Alimentos. Disponível em: <[https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca%20de%2028%2C6%25%20da,e%20Estat%C3%ADstica%20\(IBGE\)%20%E2%80%93%20](https://www.investe.sp.gov.br/setores-de-negocios/alimentos/#:~:text=Cerca%20de%2028%2C6%25%20da,e%20Estat%C3%ADstica%20(IBGE)%20%E2%80%93%20)>. Acesso em: 25/10/2023.

Amazon S3 Documentation:

Amazon Web Services. Amazon Simple Storage Service (S3) Documentation. Disponível em: <https://docs.aws.amazon.com/s3/>. Acesso em: 13/11/2023.

Boto3 Documentation:

Amazon Web Services. Boto3 Documentation. Disponível em: <https://boto3.amazonaws.com/v1/documentation/api/latest/index.html>. Acesso em: 13/11/2023.

Barocas, S., & Hardt, M. (2019). Justiça e Abstração em Sistemas Sociotécnicos. Em Anais da Conferência sobre Justiça, Responsabilidade e Transparência (FAT/ML 2019).

Brasil. (2018). Lei Nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados. Recuperado de http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm

Crawford, K. (2017). Um Algoritmo Pode Ser Agônico? Dez Cenas da Vida em Públicos Calculados. *Science, Technology, & Human Values*, 42(1), 77–92.

Diakopoulos, N. (2016). A Ética dos Algoritmos: Mapeando o Debate. *Big Data & Society*, 3(2), 2053951716679679.

Dignum, V., & outros. (2018). Inteligência Artificial Responsável: Governando a IA e os Desafios Futuros.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Sartor, G. (2018). AI4People—Um Framework Ético para uma Boa Sociedade de IA: Oportunidades, Riscos, Princípios e Recomendações. *Minds and Machines*, 28(4), 689–707.

Hardt, M., Price, E., & Srebro, N. (2016). Igualdade de Oportunidades em Aprendizado Supervisionado. Em Avanços em Sistemas de Informação Neural 29 (NIPS 2016).

Jones, H. E., York, W. B., & Walton, R. N. (2020). Os Desafios da Ética em Big Data. *The Harvard Data Science Review*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). A Ética dos Algoritmos: Um Mapa. *Journal of Data Protection & Privacy*, 1(1), 30–46.

O'Neil, C. (2016). Armas de Destrução Matemática: Como o Big Data Aumenta a Desigualdade e Ameaça a Democracia. Crown.

Nações Unidas. (2015). Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável. Recuperado de <https://sdgs.un.org/2030agenda>

15. Anexos