Integration

BIG DATA

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA

INTEGRATION

Autores:

Gabriel Santos

Henri Harari

Rafael Moritz

Patrick Miranda

Raduan Muarrek

Vitor Moura

Data de criação: 24 de Outubro de 2023

SÃO PAULO – SP

2023

Controle de Documento

Histórico de Revisões

Data	Autor	Versão	Resumo da atividade
26/10/2023	Patrick	1.0	Adição dos artefatos da sprint 1

Table 1: Controle de documento

Sumário

1. Introdução

1.1 Parceiro de Negócios

1.2 Problema

1.2.1 Definição do Problema

2. Objetivos

2.1 Objetivos Gerais

2.2 Objetivos Específicos

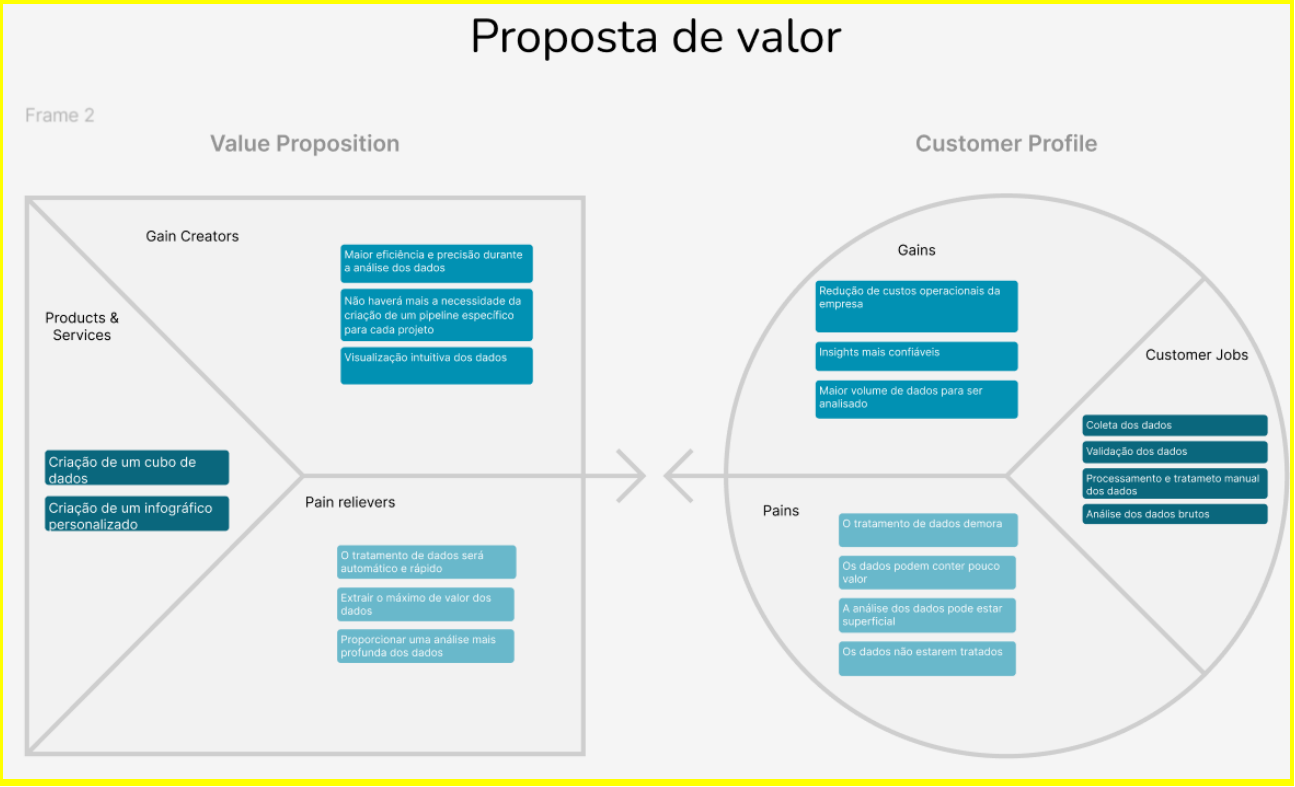
2.3 Justificativa

3. Análise de Negócios

3.1 Proposta de Valor

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

imagem 1: Canvas proposta de valor



Fonte: Dados dos autores (2023)

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

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

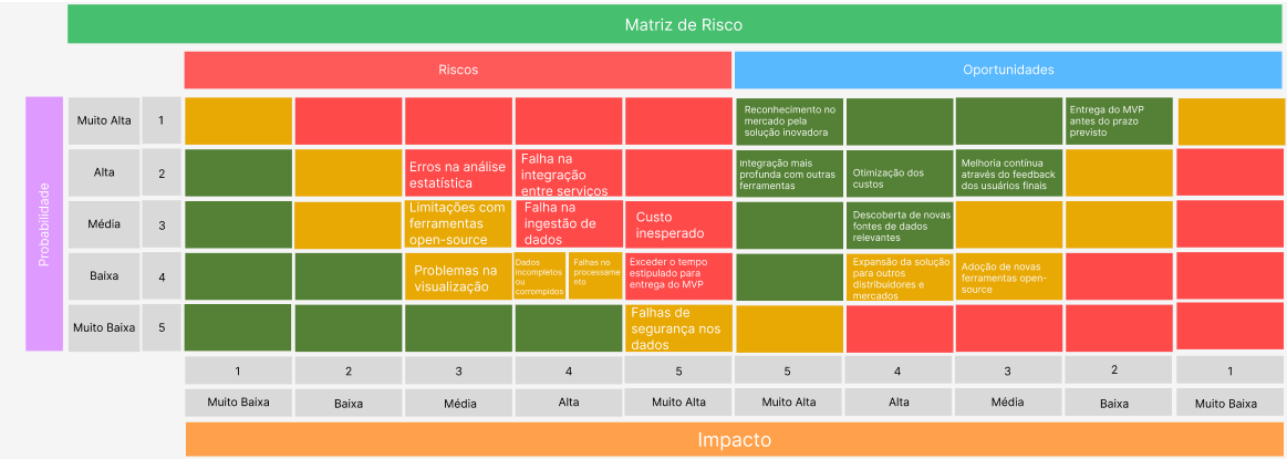
Gains (Ganhos) e Gain Creators (Criadores de Ganho):
Os "Ganhos" expressam os resultados positivos e benefícios que as consultorias desejam alcançar, como otimização de estratégias de "go to market", maior eficiência operacional e geração de insights inovadores para seus clientes na indústria alimentícia.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

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

3.2 Matriz de Risco

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

imagem 1: matriz de risco



Fonte: Dados dos autores (2023)

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

imagem 2: plano de ação

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

Fonte: Dados dos autores (2023)

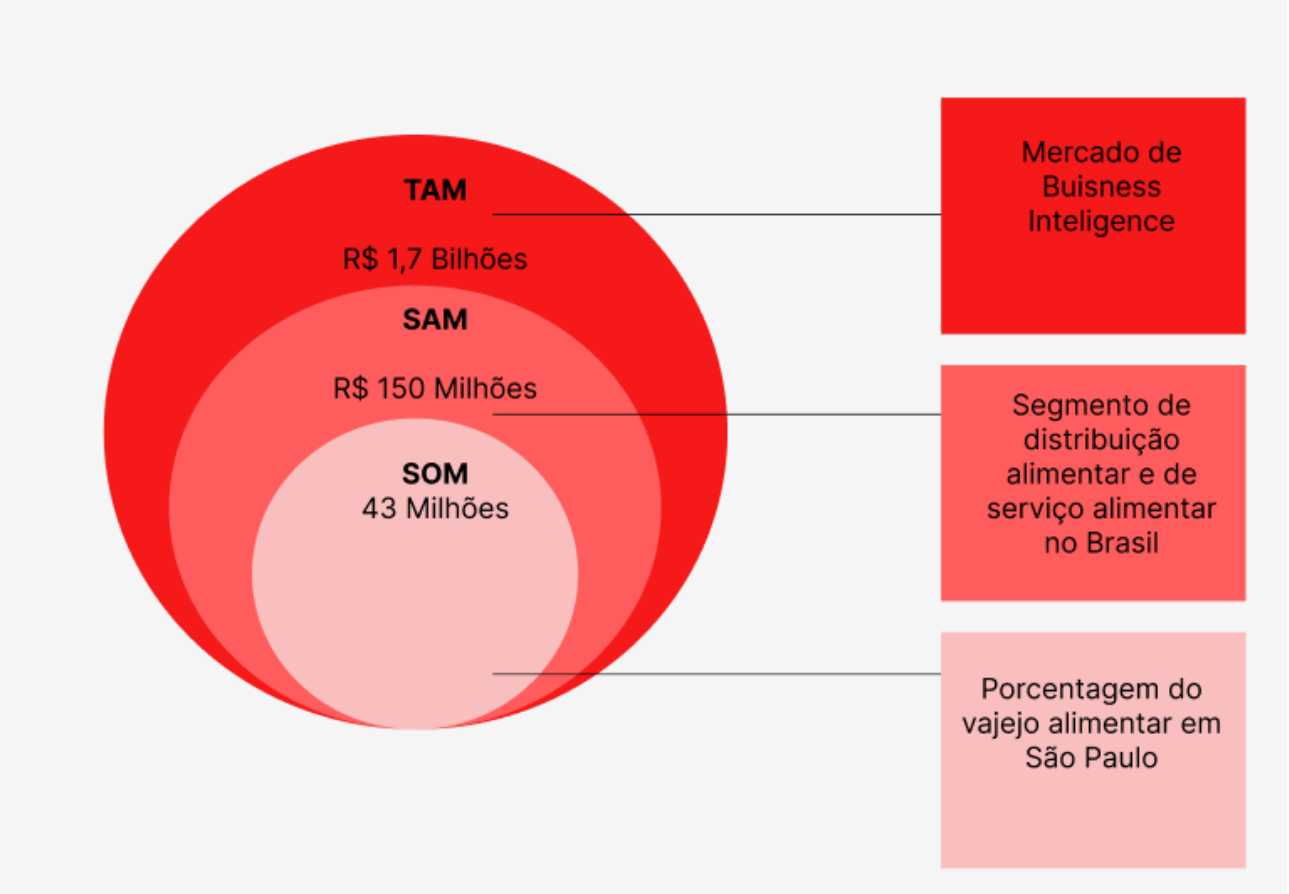
3.3 TAM SAM SOM

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

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

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

Imagem 1: TAM SAM SOM



TAM

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

Premissa: Existe uma grande demanda por insights e análises na indústria alimentícia. O TAM engloba todas as consultorias, empresas e indivíduos que poderiam potencialmente se beneficiar do uso de ferramentas de Business Intelligence na indústria alimentícia.

Valor: R$ 1,7 bilhões.

SAM

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

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

Valor: R$ 150 milhões.

SOM

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

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

Valor: R$ 43 milhões.

4. Análise de Experiência do Usuário

4.1 Personas

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

Imagem 1: Persona, Moisés Aragão

Fonte: Dados dos autores (2023)

Imagem 2: Persona, Enzo Ananias

Fonte: Dados dos autores (2023)

4.2 Jornada do Usuário

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

Imagem 1: jornada, Enzo Ananias

Fonte: Dados dos autores (2023)

Imagem 2: jornada, Moisés Aragão

Fonte: Dados dos autores (2023)

4.3 User Stories

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

Número	01
Título	Dados governamentais atualizados.
Personas	Moisés Aragão
História	Eu como teach lead, quero que seja possível atualizar os dados governamentais salvos na solução, de forma a garantir que os dados utilizados para análise reflitam a realidade do País.
Critérios de aceitação	1. Deve ser possível enviar novos dados para atualizar o banco de dados. 2. Os dados consumidos para criação do cubo de dados devem estar atualizados.
Testes de aceitação	Critério 1: - Os novos dados governamentais foram enviados para o RDS. - Aceitou: Correto, começar o próximo critério - Recusou: Errado, analisar formato e origem dos dados - Dados indevidos foram enviados para o RDS: - Aceitou: Errado, revisar configuração do RDS - Recusou: Correto, começar o próximo critério Critério 2: - Os dados consumidos correspondem aos novos dados enviados. - Aceitou: Correto, começar o próximo critério - Recusou: Errado, revisar código de consumo do RDS - Os dados consumidos não representam os dados atualizados: - Aceitou: Errado, revisar configuração de consumo do Rds - Recusou: Correto, começar o próximo critério

Número	02
Título	Visualização do infográfico
Personas	Enzo Ananias
História	Eu, como consultor de marketing, quero ser capaz de visualizar o infográfico gerado com base no cubo de dados, para que eu possa obter insights de qualidade com maior facilidade.
	1. Usuários com permissão devem ser capazes de visualizar o infográfico.

Critérios de aceitação	

Testes de aceitação	Critério 1:
	- Usuário autorizado foi capaz de visualizar o infográfico
	  - Aceitou: Correto, começar o próximo critério
	  - Recusou: Errado, analisar origem do erro e resolvê-lo
	- Usuário não autorizado foi capaz de visualizar o infográfico
	  - Aceitou: Errado, analisar origem do erro e resolvê-lo
	  - Recusou: Correto, começar o próximo critério

Número	03
Título	Dados do Cliente atualizados.
Personas	Moisés Aragão
História	Eu como teach lead, quero que os dados do cliente sejam semanalmente atualizados, de forma a garantir que os dados analisados do cliente estejam sempre atualizados.
Critérios de aceitação	1. Os dados consumidos pela Api do parceiro devem estar atualizados. 2. O consumo deve estar sendo realizado semanalmente.

Testes de aceitação	Critério 1:
	- Verificar se os dados da API do parceiro estão atualizados.
	  - Estão: Correto, começar o próximo critério
	  - Não estão: Errado, entrar em contato com o parceiro para solicitar revisão dos dados fornecidos
	Critério 2:
	- Verificar se o consumo ocorreu na data programada.
	  - Ocorreu: Correto, começar o próximo critério
	  - Não ocorreu: Errado, revisar código de consumo da api do parceiro

Número	04
Título	Atualização do infográfico.
Personas	Enzo Ananias
História	Eu como consultor de marketing, quero que o infográfico reflita a realidade do cubo de dados, de forma a garantir que os dados visualizados reflitam a devida realidade do parceiro.
Critérios de aceitação	1. Os dados consumidos do cubo devem condizer com as informações exibidas no infográfico. 2. Os dados consumidos devem ser os mais recentes gerados.
Testes de aceitação	Critério 1: - Verificar se as informações no cubo e nos infográficos correspondem. - consistentes: Correto, começar o próximo critério - inconsistentes: Errado, revisar código de consumo dos dados para geração dos infográficos Critério 2: - Verificar se os infográficos foram devidamente atualizados junto ao cubo de dados. - Dados recentes: Correto, começar o próximo critério - Dados anteriores: Errado, revisar código de consumo dos dados para geração dos infográficos

5. Análise Exploratória

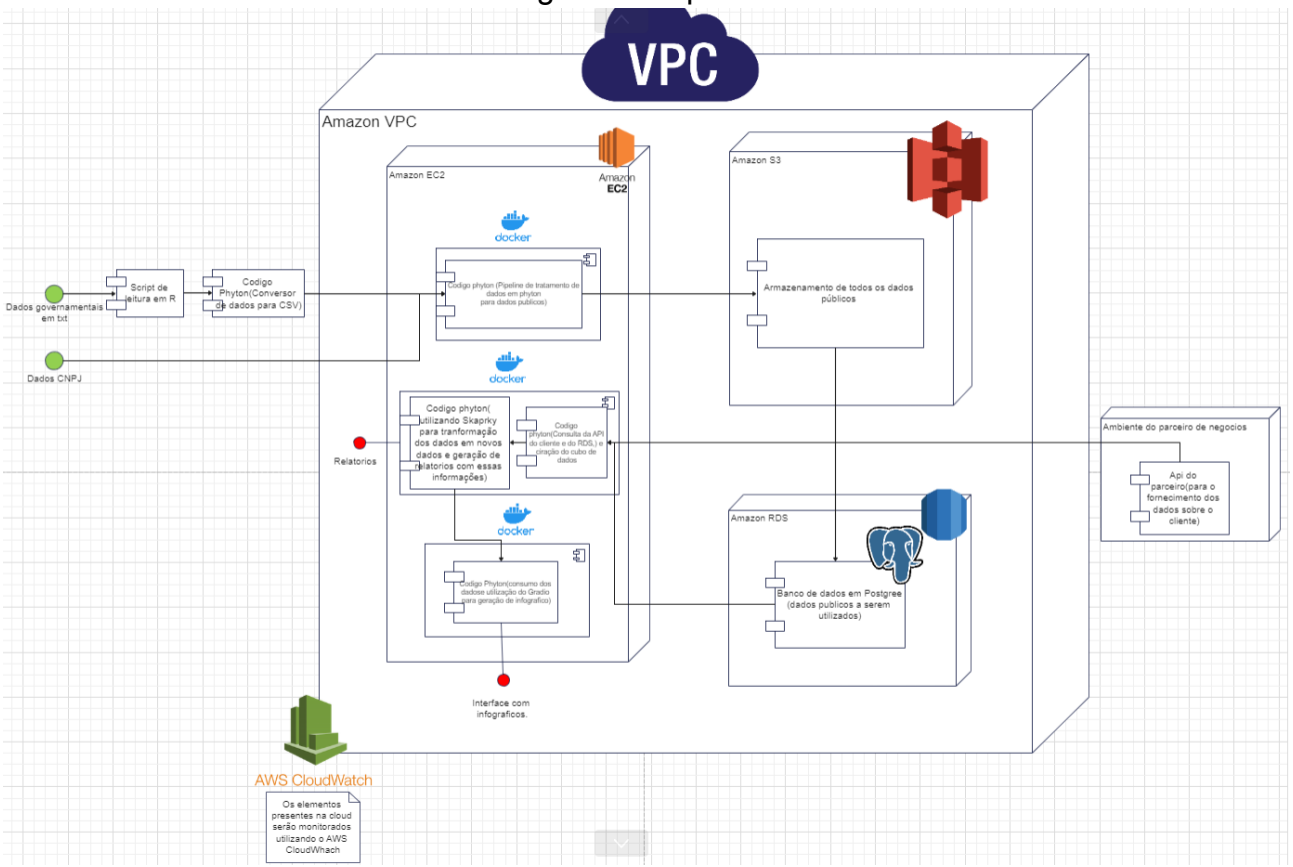
5.1 Introdução

5.2 Método

6. Arquitetura Macro

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

Imagem 1: Arquitetura



Fonte: Dados dos autores (2023)

Abaixo se encontra uma análise da arquitetura construída:

6.1 Identificação dos dados:

6.1.1 Dados públicos:

São fornecidos pelos canais de pesquisa do governo, tendo como mais frequente o IBGE, através de arquivos .txt em uma pasta no início do projeto. De partida, os dados não apresentavam-se estruturados, sendo fornecido um script em R para leitura desses dados e a criação de tabelas que possibilitam a compreensão. Após a leitura os dados

serão convertidos para .rds e posteriormente para .csv e passados para o EC2. Nos dados públicos tambem se encaixam os dados referentes aos CNPJ no Brasil, que foram forncecidos diretamente em .csv.

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

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

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

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

Serviço não monetário - POF2	Pesquisa sobre serviços não monetários na Pesquisa de Orçamento Familiar (POF) 2.
Serviço não monetário - POF4	Pesquisa sobre serviços não monetários na Pesquisa de Orçamento Familiar (POF) 4.

6.1.2 Dados do cliente:

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

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

6.2 Gestão de dados:

6.2.1 Dados públicos:

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

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

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

6.2.2 Dados privados:

Já para os dados privados os dados serão ingeridos a partir da Api que fornece os dados do cliente em questão para cada projeto. A Api será consultada por um código python rodando no EC2, e posteriormente carregado, em um código python, também na EC2, utilizando sparky para uma manipulação dos dados. Eles também serão utilizados para a criação de um cubo de dados.

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

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

6.3 Seleção dos serviços AWS:

Tratamento/processamento de dados:

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

Serão no total quatro códigos alocados na EC2:

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

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

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

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

Armazenamento:

S3: Por aspectos do pipeline de big data como volume e custos, será utilizado o , serviço que permite o armazenamento de dados de forma flexível e não estruturada. Nele se espera comportar todos os dados públicos com alto volume histórico de datação prezando pela economia em espaço e processamento alcançáveis por uma base de dados não relacional.

RDS: Posteriormente, após a seleção dos dados de interesse para as análises de mercado e negócios, será utilizado o AWS RDS para hospedar um banco de dados relacional Postgre, o qual receberá os dados de mais importância que estavam no S3. Será restrito ao último os dados desejados para a elaboração de tabelas e informações de negócios na próxima etapa de tratamento.

Segurança de dados:

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

Monitoramento:

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

6.4 Fluxo dos dados:

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

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

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

Cubo de Dados: Após a formação inicial do cubo de dados os dados passarão por outro código python no EC2, sendo processados e agregados de forma a gerar as informações

que o cliente deseja para gerar um relatório. E por fim serão enviados para outro código python no EC2, o qual irá gerar infográficos a partir desses dados.

6.5 Segurança:

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

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

6.6 Monitoramento e gerenciamento:

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

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

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

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

7. Conclusões

8. Referências

9. Anexos