



igti

RELATÓRIO

PROJETO APLICADO

Instituto de Gestão e Tecnologia da Informação
Relatório do Projeto Aplicado

Disponibilização de dados analíticos e recomendações para clientes no Varejo E-commerce

Gabriel Lima Novais

Orientador(a):
Murillo Barbosa

14/01/23



GABRIEL LIMA NOVAIS

INSTITUTO DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO

RELATÓRIO DO PROJETO APLICADO

Disponibilização de dados analíticos e recomendações para clientes no Varejo E-commerce

Relatório de Projeto Aplicado
desenvolvido para fins de conclusão do
curso MBA em Engenharia de Dados.

Orientador (a): Murillo Barbosa

Niterói
14/01/23

Sumário

1. CANVAS do Projeto Aplicado	4
1.1 Desafio	5
1.1.1 Análise de Contexto	5
1.1.2 Personas	9
1.1.3 Benefícios e Justificativas	12
1.1.4 Hipóteses	14
1.2 Solução	17
1.2.1 Objetivo SMART	17
1.2.2 Premissas e Restrições	18
1.2.3 Backlog de Produto	21
2. Área de Experimentação	24
2.1 Sprint 1	24
2.1.1 Solução	24
• Evidência do planejamento:	24
• Evidência da execução de cada requisito:	27
• Evidência dos resultados:	32
2.1.2 Experiências vivenciadas	34
2.2 Sprint 2	35
2.2.1 Solução	35
• Evidência do planejamento:	35
• Evidência da execução de cada requisito:	35
• Evidência dos resultados:	35
2.2.2 Experiências vivenciadas	35
2.3 Sprint 3	36
2.3.1 Solução	36
• Evidência do planejamento:	36
• Evidência da execução de cada requisito:	36
• Evidência dos resultados:	36
2.3.2 Experiências vivenciadas	36
3. Considerações Finais	37
3.1 Resultados	37
3.2 Contribuições	37
3.3 Próximos passos	37

1. CANVAS do Projeto Aplicado

Segue a Figura 01 conceitual canvas do projeto aplicado, que representa todas as etapas do projeto.



Figura 01 - Canvas do Projeto Aplicado

1.1 Desafio

1.1.1 Análise de Contexto

O varejo, principalmente em países em desenvolvimento como o Brasil, é de suma importância, por ser uma fonte provedora de empregos. No Brasil corresponde cerca de 47,4% do PIB, apresentando um crescimento expressivo e sempre está na mira de empreendedores em busca de oportunidades.

O setor consiste no processo de compra de produtos em quantidades relativamente grande dos produtores atacadistas e outros fornecedores e posteriormente vendidos em quantidades menores ao consumidor final. Além disso, pode ser considerado como atividade responsável por providenciar mercadorias e serviços desejados pelos consumidores.

O varejo é definido, segundo Philip Kotler como método comercial que engloba todas as atividades relativas à venda direta de produtos ou serviços ao consumidor final. No entanto, Parente (2000), conceitua o varejo como todas as atividades que englobam o processo de venda de produtos e serviços para atender a uma necessidade pessoal do consumidor final. Las Casas (2004, p. 17) cita uma definição de varejo de acordo a American Marketing Association, onde considera o varejo como “uma unidade de negócio que compra mercadorias de fabricantes, atacadistas e outros distribuidores e vende diretamente a consumidores finais e eventualmente aos outros consumidores”.

Então qualquer empresa que forneça um produto ou serviço para o consumidor final está praticando varejo. O comércio varejista conta com processos sistematizados, a fim de atender bem às demandas dos clientes. É importante apresentar uma cadeia de suprimentos bem definida para corresponder às necessidades do mercado e demonstrar competitividade. O comércio varejista vem assumindo uma importância crescente no panorama empresarial do Brasil e do mundo.

À medida que as empresas varejistas se expandem, passam a adotar tecnologias avançadas de informação e de gestão e desempenham papel cada

vez mais importante na modernização do sistema de distribuição e da economia brasileira. Para que o setor seja bem-sucedido, é preciso estar alerta e pronto para se adaptar aos desafios de mercado.

A tecnologia da informação desempenha papel primordial nesse processo de evolução, visto que o comércio virtual (e-commerce) é a grande tendência do setor de varejo. O varejo online fornece ao consumidor informações, agilidade na entrega e apresenta descontos atraentes. Não apenas a tecnologia servirá como meio que possibilita a realização da venda, através das plataformas digitais, mas como uma forma essencial de permitir análises de planejamento estratégico, marketing, performance e muitas outras de acentuada importância para tomadas de decisão.

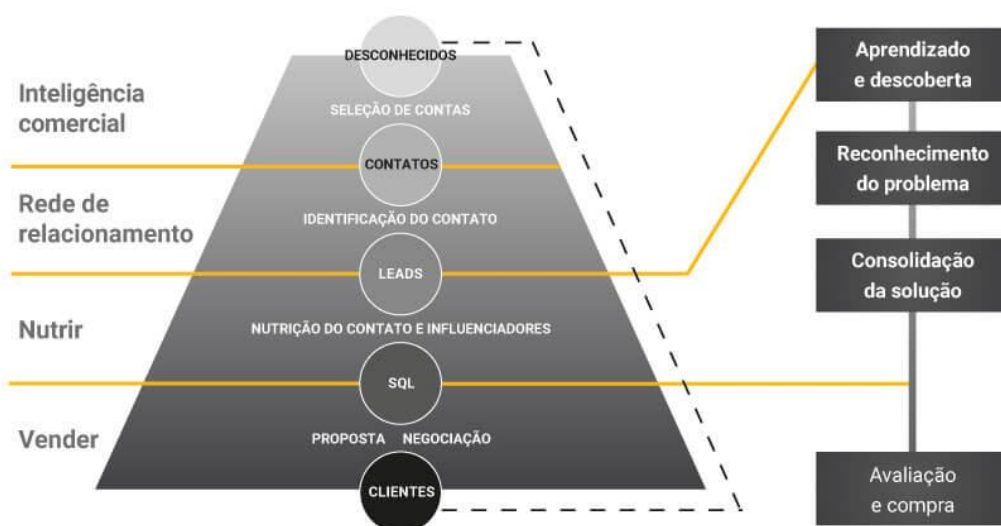


Figura 02 - Account Based Marketing

Fonte: <https://layerup.com.br/account-based-marketing/>

Segundo afirma a Associação Brasileira de Comércio Eletrônico, as vendas online correspondem a 11,6% do setor varejista do país. Contudo, toda esta evolução causa um impacto no comportamento do consumidor.

A tendência do consumidor moderno é ser mais exigente, mais direcionado por questões de tempo, mais necessitado de informação e

altamente individualista. Ações que visam a construção de um ambiente de vendas mais personalizado e com um conforto maior para o consumidor, garantem elevação de rentabilidade ao negócio.



Figura 03 - Modelo de Matriz CSD

Dessa forma, após elucidação do contexto e da exposição do problema, é possível defender cada vez mais o uso intensivo de dados. A necessidade de se construir maneiras mais eficientes de garantir uma experiência do usuário, ao mesmo tempo que se permite um ferramental adequado para acompanhar a performance das vendas, preços e visitas do site e dos desdobramentos de tais monitorias, se baseiam quase que completamente nessa hipótese.

Para estruturar as idéias mencionadas foram realizadas a matriz e a observação contida na Figura 03 acima. Adicionalmente, para realizar a análise do contexto do problema, foi utilizado o canvas abaixo, contido na Figura 04.

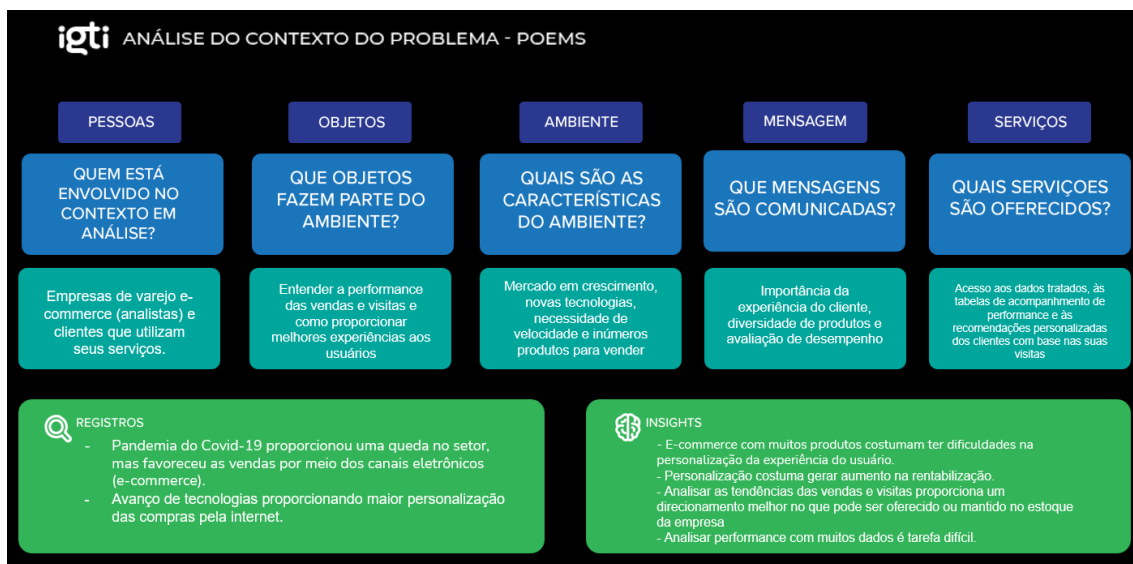


Figura 04 - Análise do Contexto do problema

No desenvolvimento de uma solução, um comum exercício que busca identificar o perfil daqueles que serão os usuários é a criação de personas, não necessariamente representando alguém do público alvo que saiba tudo do negócio em questão. Especificamente para este projeto, foram utilizadas as informações detalhadas nas seções anteriores para criar 3 perfis de personas. Esses perfis são de duas “pontas” completamente distintas do negócio em questão.

Enquanto a primeira e a segunda persona que são funcionários da empresa, do local onde a solução será empreendida, sendo os dois analistas de áreas diferentes, a terceira persona representa o consumidor final dos serviços apresentados. De forma resumida, listam-se essas personas:

- Analistas de BI e de Marketing dentro da empresa, na qual a solução será utilizada como um guia para tomada de decisão.
- Clientes, consumidores dos produtos vendidos pelo e-commerce, que desfrutarão de uma experiência personalizada de compra na plataforma.

Persona 1

Nome: Pedro

Idade: 30 anos

Profissão: Analista de BI

Características comportamentais: Pedro é engenheiro de produção e possui uma formação mais analítica. Possui um perfil bem pragmático, gosta de resolver problemas gerando formas de solução que funcionem no dia a dia, não precisando recorrer a complexidades desnecessárias. Ele precisa de dados para acompanhar de maneira mais eficiente as métricas de vendas e visitas da empresa e entender como esses indicadores refletem o valor gerado para companhia. Ele deseja rapidez e facilidade para gerar os reports e visualizações de dados para garantir tomadas de decisões mais concretas sobre preços e atingimento de metas.

Persona 2

Nome: Maria

Idade: 28 anos

Profissão: Analista de Marketing

Características comportamentais: Maria é formada em administração e possui pós-graduação em marketing digital. Ela possui um perfil mais criativo, com boas análises sobre tendências de mercado e conhecimentos sobre ações de marketing e segmentação de audiências, o que lhe confere resultados bons nas suas campanhas de marketing. Ela possui uma dificuldade de entender como melhorar os produtos recomendados para seus clientes e como segmentá-los para melhorar suas campanhas de push (envio de mensagens e acionamento de clientes).

Persona 3

Nome: Sérgio

Idade: 54 anos

Profissão: Carpinteiro

Características comportamentais: Sérgio possui curso técnico em carpintaria, e no seu negócio, prioriza pela qualidade do acabamento dos seus móveis. Entretanto precisa de ferramentas e outros insumos. Como possui muitas demandas, não pode perder tempo procurando esses recursos e para isso realiza compras online. Além disso, gosta de pescar e fazer trilhas, como forma de lazer, e compra todas as suas iscas e materiais pela internet. Se ele tiver o que precisa e gosta, em um só lugar, com certeza compraria e otimizaria seu tempo tão valioso.

As personas listadas foram construídas após feito o mapa da empatia mostrado na Figura 05 . Note que nesse mapa as dores mostradas são compartilhadas pelas 3 personas e a solução a ser proposta visa sanar essa dor. De modo que, é criada a seguinte relação lógica: Uma vez disponibilizado dados para consulta sobre indicadores e métricas de avaliação do negócio e ao mesmo tempo recomendações de produtos, os usuários terão uma experiência personalizada de consumo, aumentando a rentabilização que poderá ser

acompanhada nos dados da outra tabela, permitindo que os analistas tomem decisões estratégicas para aproveitar não apenas essa nova maneira de exposição de conteúdo ou de seleção de portfólio de produtos a serem vendidos, como aperfeiçoar as campanhas de marketing e melhor interpretar as audiências e seus perfis.

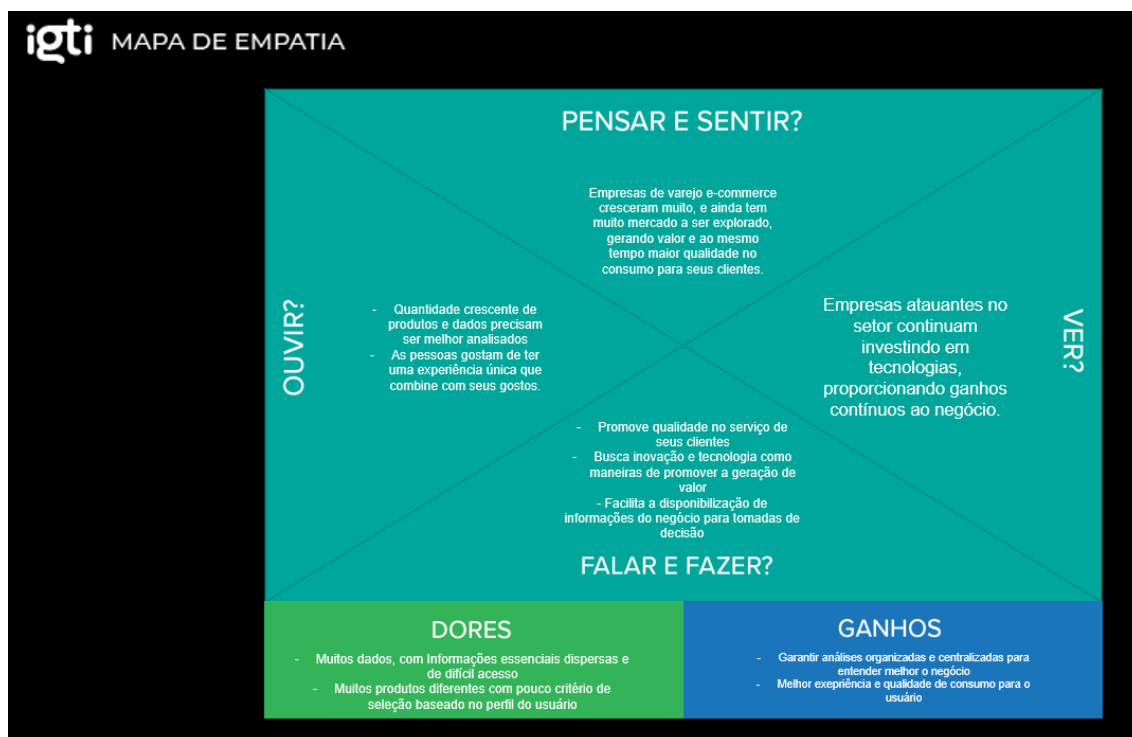


Figura 05 - Mapa de Empatia

1.1.3 Benefícios e Justificativas

Conforme apresentado nas seções anteriores, o problema está bem delineado, ou seja, conforme o comércio varejista se expandiu em especial por conta de fatores externos, como a pandemia e evolução tecnológica, o volume de dados, produtos e serviços se expandiram também. Para que seja possível melhorar o acompanhamento e avaliação de métricas e operações com a finalidade de proporcionar tomadas de decisões assertivas, é necessário intensificar o uso de tecnologias relacionadas a big data, processamento e armazenamento de dados em cloud e afins.

O objetivo do projeto é o de proporcionar o tratamento de dados, armazenamento de dados processados e disponibilização de dados para duas finalidades distintas: uma focada nas análises de suporte de tomada de decisão, e outra com a finalidade de melhorar a personalização da experiência do usuário.

Os benefícios auferidos dessa empreitada podem ser observados nas facilidades de obtenção de dados, que estarão disponibilizados em formato de tabela SQL, contendo métricas de avaliação, que inclusive dada a natureza da infraestrutura aplicada, permite expansão de funcionalidades e adição de colunas. Além desses benefícios, existirão dados também disponibilizados em bases que permitem consultas em SQL, que fornecem a reelações de recomendações de produtos e usuários que facilitarão o uso de vitrines, ações de acionamento de clientes, segmentação de audiências e outras finalidades (inclusive mensuradas pelos dados da outra tabela)

O Blue print (Figura 06), divide a experiência em ação, funcionalidade, interação e mensagem. Deve-se entender a ação da persona ao buscar uma solução da sua dor. Para cada ação, a persona verá funcionalidades para as soluções possíveis, como interagir com as possíveis soluções e qual mensagem ela absorve durante aquela ação.



Figura 06 - Blueprint

A ferramenta da proposição de valor (Figura 07) detalha como que a persona enxerga o valor na solução a ser desenvolvida diante das dores que enfrenta. A reflexão de como a persona irá extrair valor ajuda a tornar a proposta mais objetiva. Como resultado, obteve-se que a solução mais adequada está direcionada em fornecer uma fonte de dados que auxilia a tomada de decisão.



Figura 07 - Explicação de Proposição de Valor

1.1.4 Hipóteses

As hipóteses seguem expostas na Figura 08, nela as observações feitas pelas personas em uma coluna e na coluna ao lado as hipóteses criadas para desenvolver o produto. Com essa informação foi possível estabelecer especificações mais direcionadas e detalhadas que eventualmente serão consideradas no processo macro de desenvolvimento do projeto.

	Observações	Hipóteses
1	Pandemia do Covid-19 proporcionou uma queda no setor, mas favoreceu as vendas por meio dos canais eletrônicos (e-commerce).	Maiores quantidades de informações navegacionais e transacionais foram e continuam sendo produzidas pelas empresas de varejo. É possível utilizar essa abundância de informações para melhoria do negócio
2	Evolução do comportamento do consumidor que possui maiores exigências, gosta de escolher o conteúdo que consome, é autoral e precisa sanar suas necessidades de forma rápida e precisa. Grande migração para meios eletrônicos e redução de compras por meios físicos.	É possível conquistar o cliente e fidelizá-lo por meio de personalização da experiência, promoções diferenciadas e o entendimento de suas necessidades de maneira mais específica. Para isso precisa-se de acompanhamento e facilidade na monitoria das ações.
3	Avanço de tecnologias relacionadas ao Big Data e IA, proporcionando maior coleta de informações, decisões automatizadas, decisões inteligentes e recomendações eficientes	É possível com ferramentas de armazenamento e processamento de dados de Cloud obter recursos para garantir que as melhores práticas e técnicas estejam ao dispor das soluções que visam melhorar as tomadas de decisão e a experiência do usuário na plataforma de venda.

Figura 08 - Matriz de observações para hipóteses

Verifica-se que as observações da matriz de observações para hipóteses direcionam o projeto na busca por formas de auxiliar tomada de decisões de negócios por meio de dados e as tecnologias associadas ao uso intensivo de dados.

Para elucidar as ideias que originaram a solução do projeto final usou-se um método de priorização de ideias. A matriz de priorização de ideias trata-se de uma ferramenta simples na qual costuma-se inserir as ideias validadas na vertical e os critérios de avaliação (de acordo com as principais necessidades identificadas) na horizontal.

Uma vez construída a matriz, a análise indica a ideia mais provável de sucesso e que poderá trazer maior resultado efetivo para o projeto. Dessa maneira ela serve para fornecer uma melhor visão das ideias e apoiar o processo de tomada de decisão. As pontuações atribuídas na matriz estão de acordo com a tabela na Figura 09 abaixo.

Escala	B - Benefícios	A - Abrangência	S - Satisfação	I - Investimentos	C - Cliente	O - Operacionalidade
5	De vital importância	Total (de 70 a 100%)	Muito grande	Pouquíssimo investimento	Nenhum impacto	Muito fácil
4	Significativo	Muito grande (de 40 a 70%)	Grande	Algum investimento	Impacto pequeno	Fácil
3	Razoável	Razoável (de 20 a 40%)	Média	Médio investimento	Médio impacto	Média facilidade
2	Poucos benefícios	Pequena (de 5 a 20%)	Pequena	Alto investimento	Impacto grande	Difícil
1	Algum benefício	Muito pequena	Quase não é notada	Altíssimo investimento	Impacto muito grande no cliente	Muito difícil

Figura 09 - Matriz da priorização de ideias.

Fonte: Apostila do MBA em Engenharia de Dados IGTI.

Usou-se a classificação conhecida como BASICO para se fazer a priorização de ideias. A sigla nos remete aos seguintes pontos, que devem ser pontuados:

- B (Benefícios): quais são os benefícios para a organização caso a solução seja adotada?
- A (Abrangência): quantas pessoas (clientes internos e externos) serão beneficiadas com essa decisão?
- S (Satisfação): qual é a satisfação das pessoas com a solução a ser adotada?
- I (Investimento): qual será o investimento necessário para a aplicação da solução?
- C (Cliente): Qual o impacto que o cliente sofrerá com a mudança?
- O (Operações): O quão complexa é a implantação da solução?

	Ideia	B	A	S	I	C	O	Total
1	Disponibilizar dados do Olist como simulação dos dados internos de uma empresa de varejo em duas tabelas em SQL separadas por finalidade: uma para análise de métricas e outra com recomendações por usuário, de acordo com os produtos existentes.	5	4	3	3	4	3	22
2	Criar um Aplicativo que esteja integrado com uma Interface amigável para selecionar métricas e ajustar ações de marketing automaticamente	4	4	5	1	4	2	20
3	Integrar dados de redes sociais e do IBGE para que com duas fontes a mais, disponibilizemos os dados em SQL e com visualizações das regiões num mapa sobre os municípios mais rentáveis	3	3	4	2	4	2	18

Figura 10 - Matriz da priorização de ideias.

Antes para desenvolver as ideias da matriz acima na Figura 10, foi realizado um Brainstorm levantando os aspectos que envolvem a solução final, algumas de suas características e as fontes de informação. A Figura 11 mostra um conjunto de ideias separadas por finalidade e contexto. Nessa etapa considera-se que soluções coletivas trazem resultados melhores e mais criativos do que ações individuais e isoladas.

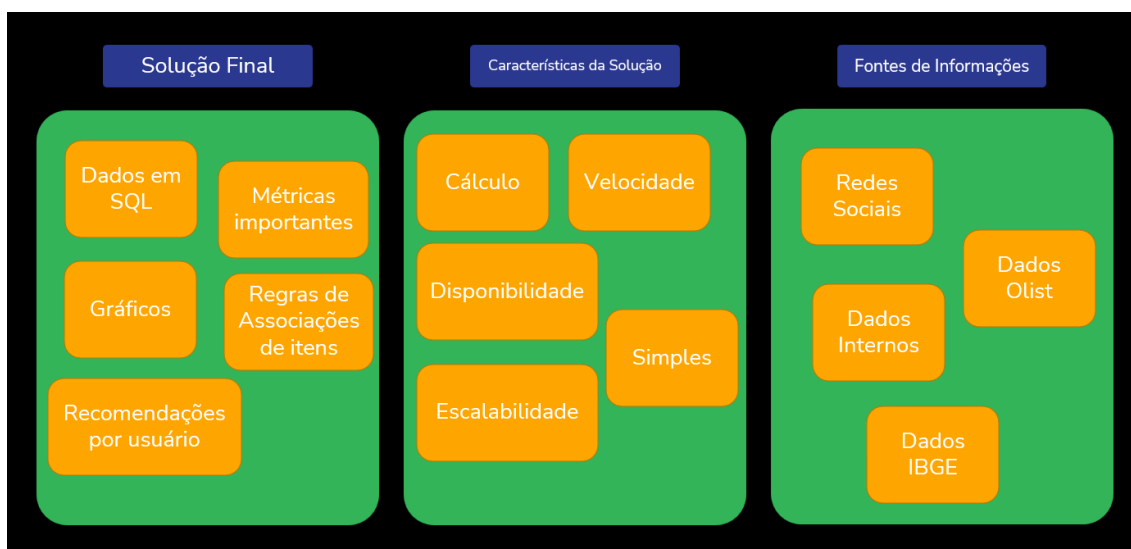


Figura 11 - Brainstorm

1.2 Solução

1.2.1 Objetivo SMART

“Objetivou-se fornecer dados tratados e com valor adicionado que possa estar acessível por um DataWarehouse disponível em um serviço de Cloud permitindo extrair análises de duas naturezas distintas.

A primeira natureza é um acompanhamento das métricas de desempenho do negócio. E a segunda é disponibilizar para clientes do Varejo E-commerce recomendações personalizadas. O usuário terá acesso a pelo menos uma plataforma para dar entrada a rotinas SQL para extrair informação dos dados. A maior parte da infraestrutura será construída em forma de código e estará em nuvem. O projeto deve ser concluído até Março de 2023.”

1.2.2 Premissas e Restrições

As premissas e restrições do projeto são importantes, uma vez que conhecendo os impactos gerados caso algo ocorra fora do esperado é possível buscar ações efetivas para mitigar os riscos. Dessa forma criou-se uma Matriz de Riscos, conforme apresentado na Figura 11. Os riscos identificados são os seguintes:

- Preços, quantidades vendidas e visitas difíceis de agrupar: significa dados muito separados e dispersos. Os dados que devem idealmente ser utilizados são dados internos, mas para fins de generalização da solução optou-se por utilizar dados do Olist que representam com louvor os dados comumente encontrados nas empresas de varejo e-commerce. O impacto potencial seria a redução de velocidade de consulta.
- Volume muito grande de produtos: significa que caso haja uma variedade muito grande de produtos, diversidade de portfólio elevada, os modelos que calculam as regras de associação entre os dados de compra ou visitas tendem a apresentar resultados de qualidade não tão bons.
- Custo elevado de infraestrutura: refere-se ao custo total necessário para conseguir implementar a solução na cloud da AWS. Esse fator de custo se torna um risco em especial pela necessidade de processamento computacional pelos clusters kubernetes que utilizam máquinas EC2. Uma boa medida preventiva adotada seria a utilização de máquinas spot ao invés de máquina on demand.

	Riscos Identificados	Impacto Potencial	Ações Preventivas	Ações Corretivas
1	Preços, quantidades vendidas e visitas difíceis de agrupar	Reduzir velocidade de consulta e agrupamento dos dados	Estudar e selecionar corretamente as melhores bases e tabelas	Aplicar uma infraestrutura que facilite o agrupamento de dados de diferentes fontes de maneira eficiente
2	Volume muito grande de produtos	Dificultar processamento e reduzir a qualidade das recomendações	Calcular quantidade média de produtos vistos por usuários	Aplicar corte de produtos com base na sua rentabilidade e/ou popularidade
3	Custo elevado de Infraestrutura (máquinas caras, processamento via cluster)	Inviabilizar cálculo de recomendações e/ou agrupamentos de dados	Realizar otimização de custos, preocupando-se com processamento e armazenamento razoável	Manter grande parte do projeto com infraestrutura em formato de código (IaC), facilitando a destruição e construção de recursos

Figura 11 - Matriz de Risco

A arquitetura construída para a solução, destaca-se na Figura 12 abaixo. Nela pode-se verificar que os dados serão distribuídos em 3 buckets no S3: um para ingestão, outro para processamento e outro para consumo.

Dessa forma a estrutura de processamento formada pelo cluster Kubernetes consegue organizar os inputs e outputs e garante a transferência desses dados de maneira rápida e fácil através do AWS Glue Crawler para o serviço de DataWarehouse da AWS, o Athena, local onde os analistas poderão realizar as consultas necessárias e eventuais diligências de ações para rentabilização.

A forma de construção da parte de buckets e serviços do Glue Crawler serão efetivadas por meio de estratégia IaC, conforme ferramental do Terraform. Todos os códigos, comandos, arquivos e instruções necessárias serão disponibilizadas no GitHub, conforme boas práticas de versionamento, no link a seguir: https://github.com/NovaisGabriel/MBA_project

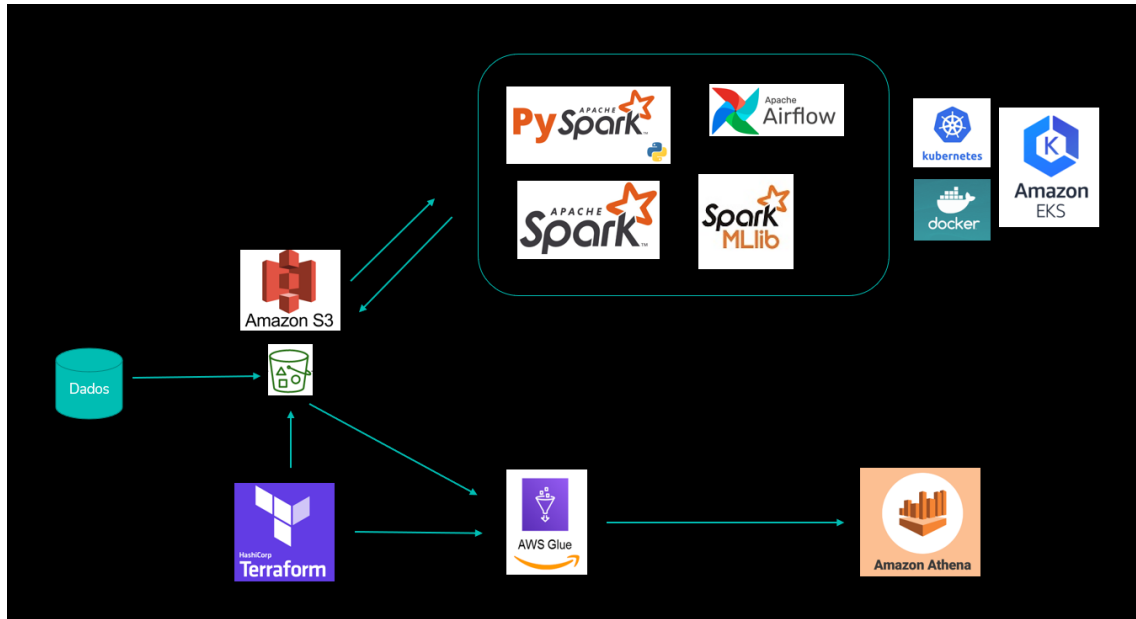


Figura 12 - Arquitetura na AWS

1.2.3 Backlog de Produto

O backlog compreende uma lista com todas as tarefas a serem desenvolvidas. As tarefas são mensuradas com prioridades, indivíduos que executarão a tarefa e instruções específicas dos requisitos necessários para que ela seja entendida como concluída. No backlog desse projeto, primeiramente, foram listados os requisitos, e conforme na Figura 13, são descritos como cada requisito será coberto ao longo dos sprints.

Na sprint 3, com sua conclusão, espera-se que a solução acompanhe um MVP, que esteja pronto e possa ser usada pelos usuários definidos. Alguns exemplos e testes serão executados para direcionar e simular a maneira como as personas do projeto extrairão o valor da solução.

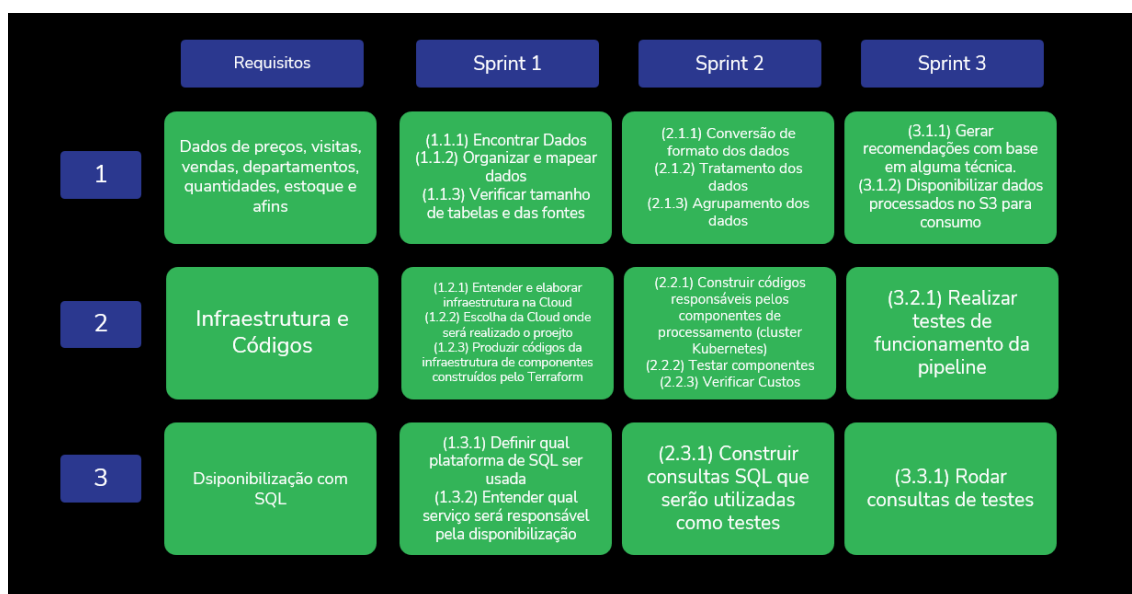


Figura 13 - Requisitos separados por sprints

A materialização do backlog descrito na Figura 13 foi registrado em ferramenta de planejamento Trello ¹. As atividades delineadas na figura acima foram organizadas nos quadros do Trello, descritos como na Figura 14. Conforme forem executadas as tarefas em andamento das Sprints do projeto, as atividades destacadas do backlog serão dadas como concluídas.

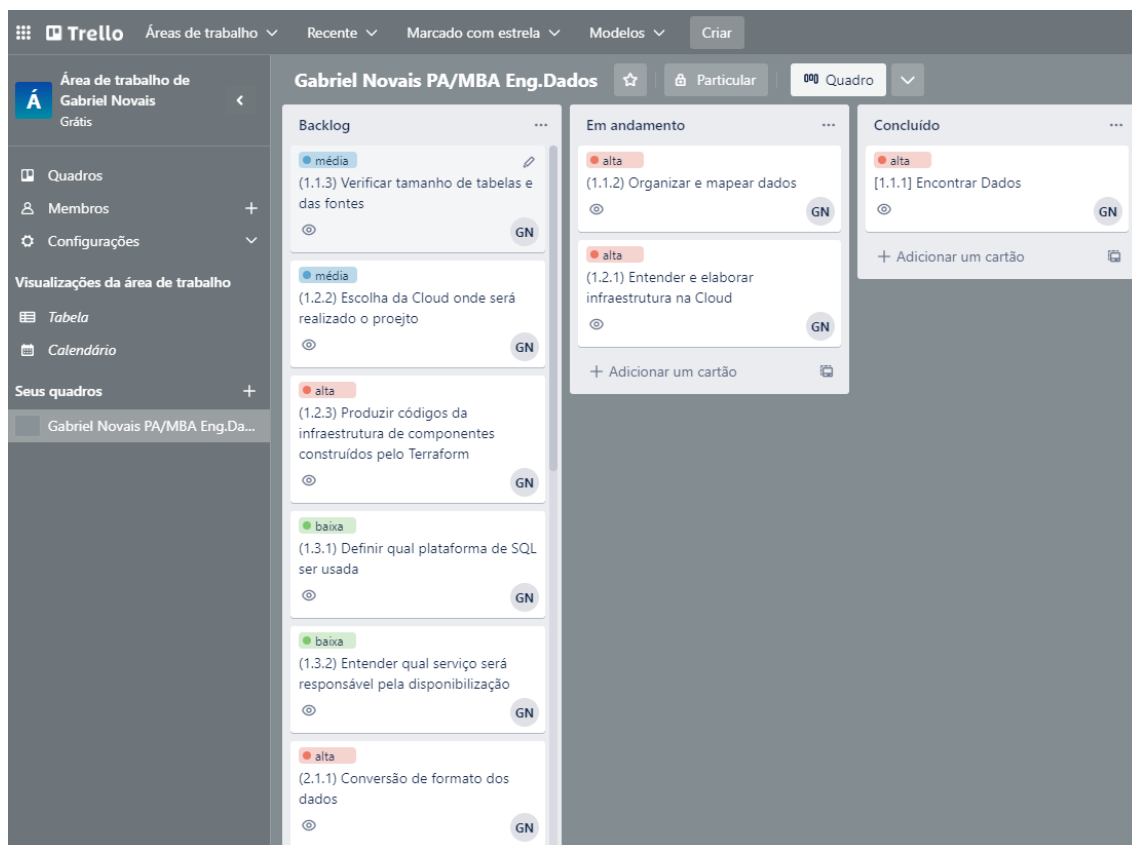


Figura 14 - Sprints, tarefas e tags no Trello

As tarefas estão numeradas conforme código, descrevendo, primeiro a Sprint, depois o requisito e terceiro a etapa. Elas são as que se seguem:

- 1.1.1.1) Encontrar Dados
- 1.1.1.2) Organizar e mapear dados
- 1.1.1.3) Verificar tamanho de tabelas e das fontes

¹ <https://trello.com/b/U8By4GoR/gabriel-novais-pa-mba-engdados>

- 1.2.1) Entender e elaborar infraestrutura na Cloud
- 1.2.2) Escolha da Cloud onde será realizado o projeto
- 1.2.3) Produzir códigos da infraestrutura de componentes construídos pelo Terraform
- 1.3.1) Definir qual plataforma de SQL ser usada
- 1.3.2) Entender qual serviço será responsável pela disponibilização
- 2.1.1) Conversão de formato dos dados
- 2.1.2) Tratamento dos dados
- 2.1.3) Agrupamento dos dados
- 2.2.1) Construir códigos responsáveis pelos componentes de processamento (cluster Kubernetes)
- 2.2.2) Testar componentes
- 2.2.3) Verificar Custos
- 2.3.1) Construir consultas SQL que serão utilizadas como testes
- 3.1.1) Gerar recomendações com base em alguma técnica.
- 3.1.2) Disponibilizar dados processados no S3 para consumo
- 3.2.1) Realizar testes de funcionamento da pipeline
- 3.3.1) Rodar consultas de testes

2. Área de Experimentação

A seção a seguir possui o objetivo de apresentar as evidências do planejamento dos requisitos selecionados do Backlog de Produto, além de mostrar a maneira como eles foram desenvolvidos e registrar os resultados alcançados. Dessa maneira é importante expor a execução e a validação dos experimentos relacionados ao desenvolvimento da solução, testando o caminho certo ou se algo precisa ser pivotado.

2.1 Sprint 1

A primeira Sprint do projeto possui como principal objetivo delinear as arquiteturas e infraestrutura necessária, além de procurar entender e elencar os objetos necessários para as atividades descritas na Sprint 2. Os dados utilizados devem também ser procurados e encontrados para a sua eventual utilização.

2.1.1 Solução

Após a elucidação dos objetivos descritos para a Sprint 1, será realizado um conjunto de evidências e descrições mais específicas das tarefas realizadas ou em progresso dessa Sprint.

- Evidência do planejamento:

1.1.1) Encontrar Dados:

Procurar dados que possuam informações sobre e-commerce que possam simular um ambiente típico enfrentado pelas empresas varejistas.

1.1.2) Organizar e mapear dados:

Entender através de um olhar mais aprofundado sobre os dados, as suas relações e dependências e mapear essas ligações.

1.1.3) Verificar tamanho de tabelas e das fontes:

No caso verificar aspectos mais quantitativos e menos qualitativos sobre a base de dados, entendendo qual seria o tamanho deles e quais tipos de dados existem naquela base.

1.2.1) Entender e elaborar infraestrutura na Cloud:

Desenhar e pensar no tipo de arquitetura que seria ideal para resolver o problema escolhido. Entender quais produtos podem auxiliar no fluxo da solução.

1.2.2) Escolha da Cloud onde será realizado o projeto:

Etapa muito relacionada à tarefa 1.2.1, pois a escolha da Cloud está muito relacionada aos tipos de produtos que serão utilizados e como eles facilitam o desenvolvimento e produção da solução.

1.2.3) Produzir códigos da infraestrutura de componentes construídos pelo Terraform:

Produzir as etapas iniciais da solução em código Terraform para providenciar os serviços básicos relacionados à solução construída.

1.3.1) Definir qual plataforma de SQL ser usada:

Etapa muito ligada aos pontos 1.2.1 e 1.2.2, nos quais definem os serviços e cloud utilizada. Nessa etapa o que se procura é definir qual o local e o serviço para que os dados possam ficar eventualmente disponíveis para consulta SQL.

1.3.2) Entender qual serviço será responsável pela disponibilização:

Etapa dedicada para entender o funcionamento da plataforma escolhida no ponto 1.3.1, no qual o que se procura é entender como ela facilita a consulta e como pode ser utilizada por analistas.

Para mostrar o progresso da Sprint segue a imagem abaixo do Trello sobre o planejamento realizado.

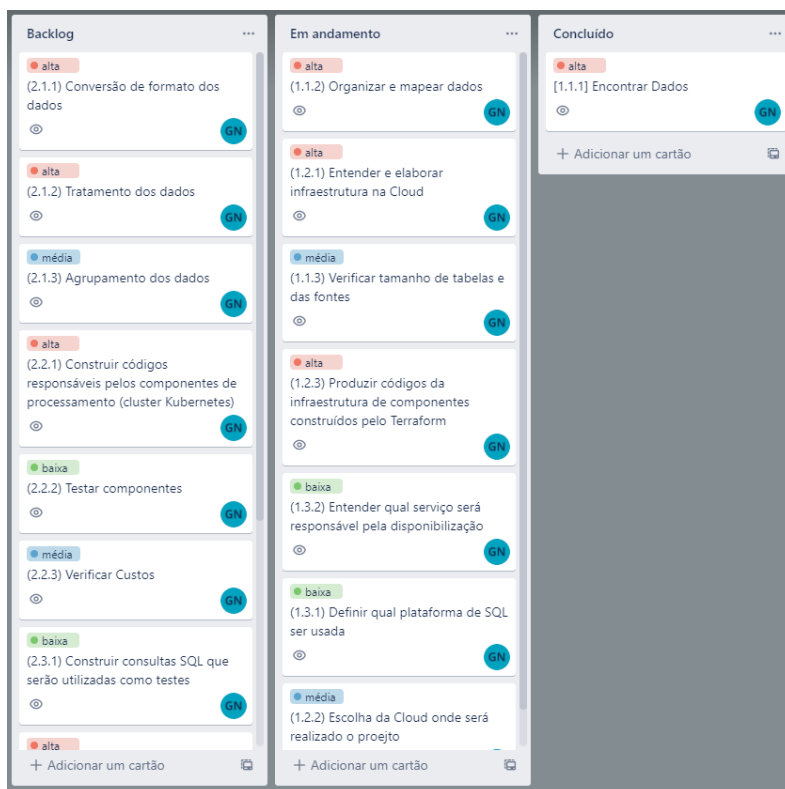


Figura 15 - Progresso da Sprint no Trello.

Após realização dos pontos destacados acima pode-se colocar cada card na coluna da direita, coluna destinada às tarefas concluídas, conforme observado na Figura 16 abaixo.

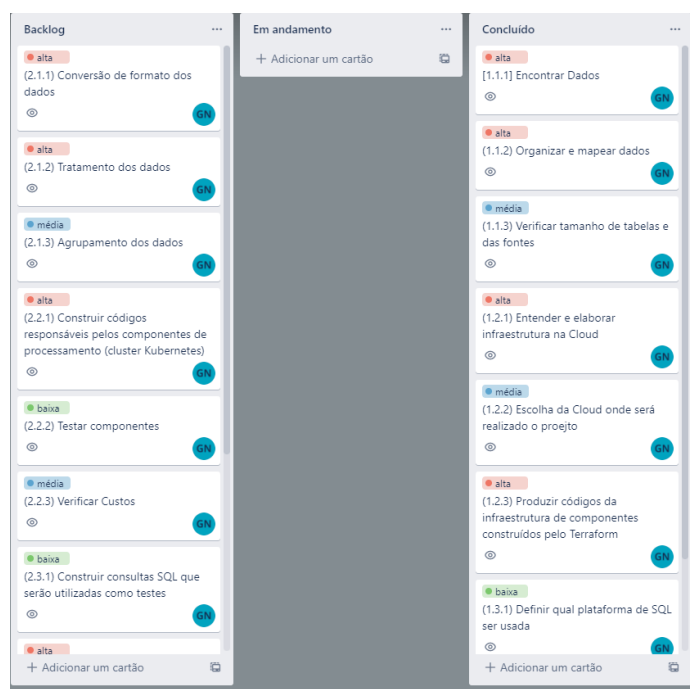


Figura 16 - Tarefas da Sprint 1 concluídas.

- Evidência da execução de cada requisito:

Após a descrição do planejamento e dos itens é possível destacar as evidências dos requisitos realizados.

Requisitos:

1.1.1) *Encontrar Dados*

1.1.2) *Organizar e mapear dados*

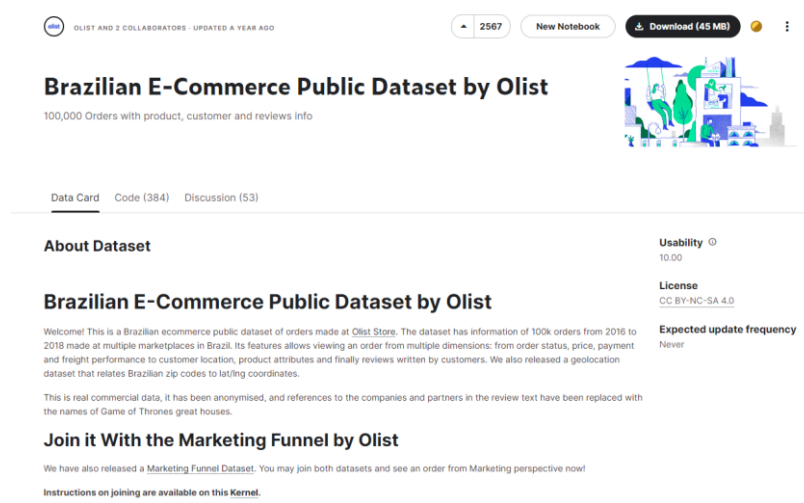
1.1.3) *Verificar tamanho de tabelas e das fontes*

Objetivos:

Procurar dados que possuam informações sobre e-commerce que possam simular um ambiente típico enfrentado pelas empresas varejistas. Entender através de um olhar mais aprofundado sobre os dados, as suas relações e dependências e mapear essas ligações. No caso verificar aspectos mais quantitativos e menos qualitativos sobre a base de dados, entendendo qual seria o tamanho deles e quais tipos de dados existem naquela base.

Evidências:

1.1.1) *Encontrar Dados*

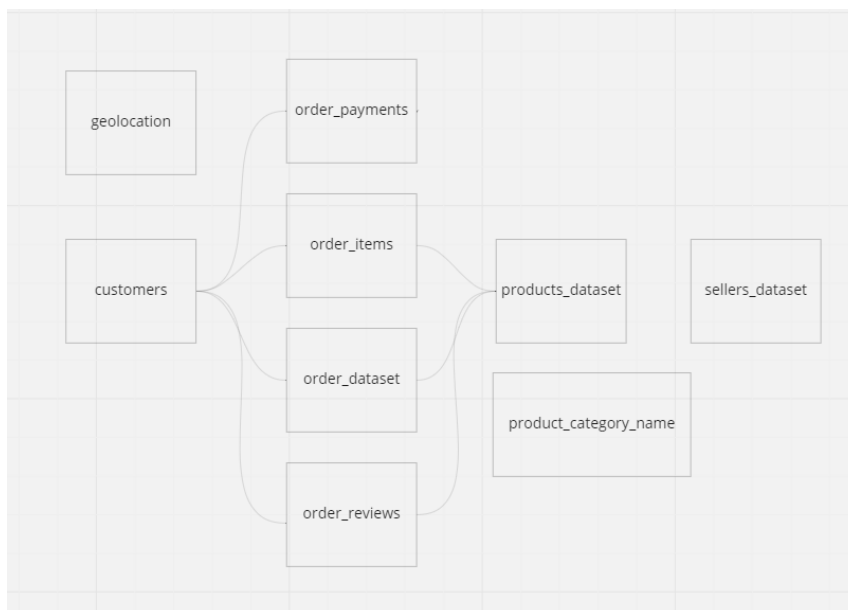


The screenshot shows the Kaggle dataset page for 'Brazilian E-Commerce Public Dataset by Olist'. At the top, it indicates 'OLIST AND 2 COLLABORATORS · UPDATED A YEAR AGO' and shows '2567' views with options for 'New Notebook' and 'Download (45 MB)'. The dataset title is 'Brazilian E-Commerce Public Dataset by Olist' with a subtitle '100,000 Orders with product, customer and reviews info'. Below the title, there are tabs for 'Data Card', 'Code (384)', and 'Discussion (53)'. The 'About Dataset' section describes the dataset as a Brazilian e-commerce public dataset of orders made at Olist Store, containing information of 100k orders from 2016 to 2018. It also mentions that the dataset has been anonymized and references to companies and partners in the review text have been replaced with the names of Game of Thrones great houses. There is a section 'Join it With the Marketing Funnel by Olist' which states that the Marketing Funnel Dataset is also available and that instructions on joining are available on this Kernel. On the right side, there is a 'Usability' section with a score of 10.00, a 'License' section with 'CC BY-NC-SA 4.0', and an 'Expected update frequency' section with 'Never'.

Link da fonte de dados:

<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

1.1.2) Organizar e mapear os dados



1.1.3) Verificar tamanho de tabelas e das fontes

Código em python para analisar formatos e tamanhos e um exemplo de análise (exemplo):

Customers

```
df = pd.read_csv("C:\\Users\\Gabriel\\Desktop\\backup\\Repositorios\\MBA_proje
print(df.head())
df.info()
```

	customer_id	customer_unique_id
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e
3	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c
4	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066

	customer_zip_code_prefix	customer_city	customer_state
0	14409	franca	SP
1	9790	sao bernardo do campo	SP
2	1151	sao paulo	SP
3	8775	mogi das cruzeiros	SP
4	13056	campinas	SP

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 5 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   customer_id                          99441 non-null  object
1   customer_unique_id                   99441 non-null  object
2   customer_zip_code_prefix             99441 non-null  int64
3   customer_city                        99441 non-null  object
4   customer_state                       99441 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.8+ MB
```

Requisitos:

1.2.1) *Entender e elaborar infraestrutura na Cloud*

1.2.2) *Escolha da Cloud onde será realizado o projeto*

1.3.1) *Definir qual plataforma de SQL ser usada*

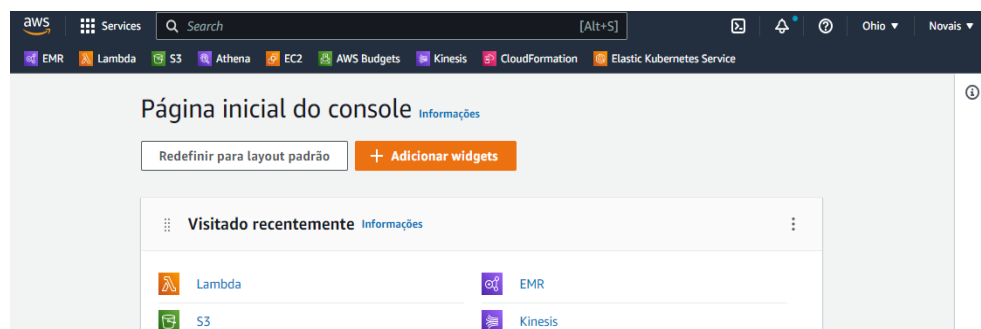
1.3.2) *Entender qual serviço será responsável pela disponibilização*

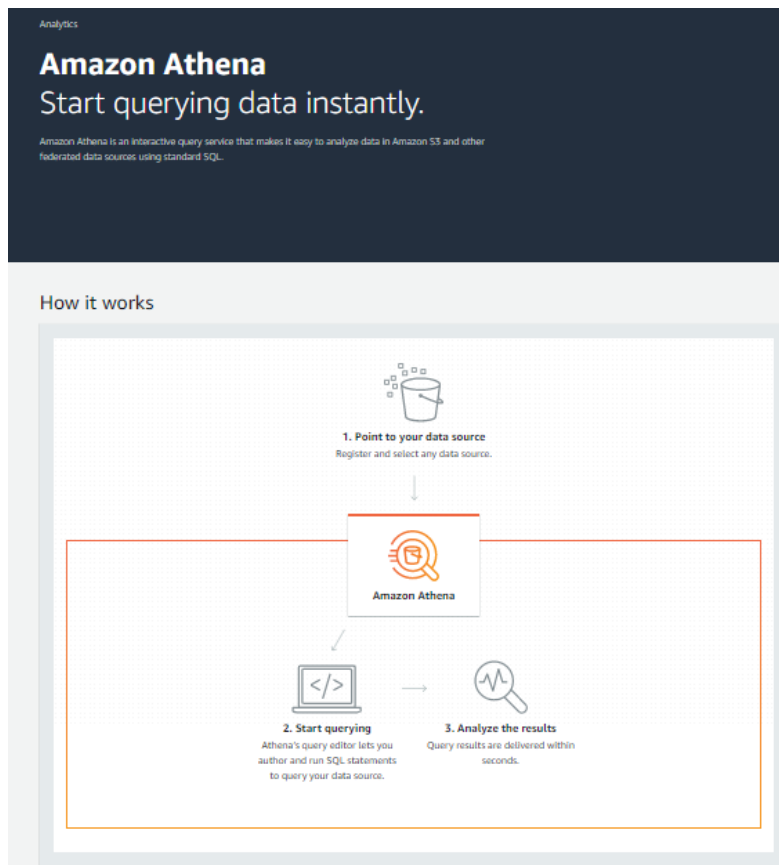
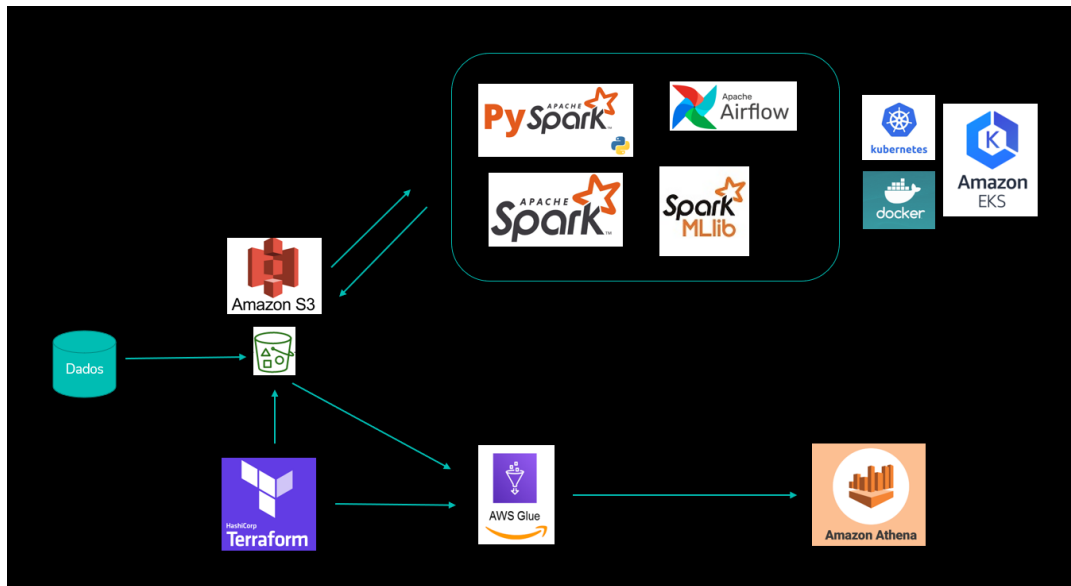
Objetivos:

Desenhar e pensar no tipo de arquitetura que seria ideal para resolver o problema escolhido. Entender quais produtos podem auxiliar no fluxo da solução. Etapa muito relacionada à tarefa 1.2.1, pois a escolha da Cloud está muito relacionada aos tipos de produtos que serão utilizados e como eles facilitam o desenvolvimento e produção da solução. Etapa muito ligada aos pontos 1.2.1 e 1.2.2, nos quais definem os serviços e cloud utilizada. Nessa etapa o que se procura é definir qual o local e o serviço para que os dados possam ficar eventualmente disponíveis para consulta SQL. Etapa dedicada para entender o funcionamento da plataforma escolhida no ponto 1.3.1, no qual o que se procura é entender como ela facilita a consulta e como pode ser utilizada por analistas.

Evidências:

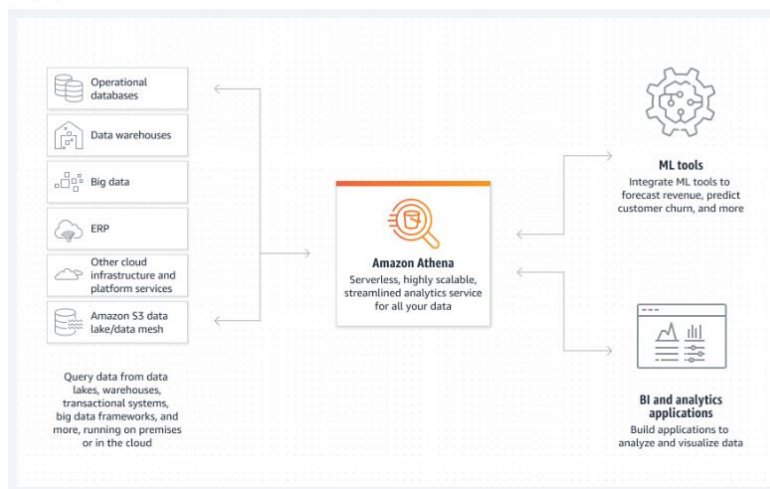
As evidências da arquitetura e da cloud escolhida podem ser vistos abaixo, de forma que cada um dos requisitos deste grupo estão sendo relacionados respectivamente aos seguintes artefatos:





Como funciona?

O Amazon Athena é um serviço de análise interativo e sem servidor criado em frameworks de código aberto, com suporte a formatos de tabela e arquivo abertos. O Athena fornece uma maneira simplificada e flexível de analisar petabytes de dados onde eles residem. Analise dados ou crie aplicações a partir de um data lake do Amazon Simple Storage Service (S3) e mais de 25 fontes de dados, incluindo fontes de dados on-premises ou outros sistemas em nuvem usando SQL ou Python. O Athena é construído com mecanismos Trino e Presto de código aberto e frameworks Apache Spark, sem necessidade de provisionamento ou configuração.



Casos de uso

Execute consultas federadas

Prepare dados para modelos de ML

Crie mecanismos distribuídos de

Analise dados do Google Analytics

Requisitos:

1.2.3) Produzir códigos da infraestrutura de componentes construídos pelo Terraform

Objetivos:

Produzir as etapas iniciais da solução em código Terraform para providenciar os serviços básicos relacionados à solução construída.

Evidências:

Os códigos necessários para construir os buckets (landing, processing e delivery), além das policies, crawlers e demais serviços necessários podem ser vistos nas imagens abaixo. Todos os códigos estão disponíveis no github.


```

infrastructure > aws > iam.tf > resource "aws_iam_role" "glue_role" > name
1 resource "aws_iam_role" "glue_role" {
2   name           = "Role_GlueCrawler"
3   path           = "/"
4   description    = "Provides write permissions to CloudWatch Logs and S3 Full Access"
5   assume_role_policy = file("./permissions/Role_GlueCrawler.json")
6 }
7
8 resource "aws_iam_policy" "glue_policy" {
9   name           = "Policy_GlueCrawler"
10  path           = "/"
11  description    = "Provides write permissions to CloudWatch Logs and S3 Full Access"
12  policy        = file("./permissions/Policy_GlueCrawler.json")
13 }
14
15 resource "aws_iam_role_policy_attachment" "glue_attach" {
16   role       = aws_iam_role.glue_role.name
17   policy_arn = aws_iam_policy.glue_policy.arn
18 }
19
20 resource "aws_iam_role" "lambda_decompress" {
21   name           = "Role_Lambda_decompress_S3"
22   path           = "/"
23   description    = "Provides write permissions to CloudWatch Logs and S3 Full Access"
24   assume_role_policy = file("./permissions/Role_Lambda_decompress_S3.json")
25 }
26
27 resource "aws_iam_policy" "lambda_decompress" {

```

- Evidência dos resultados:

Através do que foi realizado nas tarefas acima, é possível verificar que os componentes e serviços desejados foram construídos com os códigos e organizações elaboradas, de forma que os resultados são evidenciados segundo as imagens abaixo, onde pode-se ver as estruturas dos buckets, os crawlers construídos e uma query de teste no conjunto de dados da tabela costumers na landing zone.

Amazon S3 > Buckets

► Snapshot da conta
O Storage Lens fornece visibilidade sobre o uso e as tendências de atividades. [Saiba mais](#)

Buckets (6) [Info](#)
Os buckets são contêineres para dados armazenados no S3. [Saiba mais](#)

🔍 *Encontrar buckets por nome*

	Nome	Região da AWS
<input type="radio"/>	delivery-zone-715036709715	Leste dos EUA (Ohio) us-east-2
<input type="radio"/>	landing-zone-715036709715	Leste dos EUA (Ohio) us-east-2
<input type="radio"/>	processing-zone-715036709715	Leste dos EUA (Ohio) us-east-2
<input type="radio"/>	temp-functions-rony-715036709715	Leste dos EUA (Ohio) us-east-2
<input type="radio"/>	terraform-logs-gabriel	Leste dos EUA (Ohio) us-east-2
<input type="radio"/>	testekuberneteslogs	Leste dos EUA (Ohio) us-east-2

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates

Crawlers (3) Info

View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run
<input type="checkbox"/>	dl_delivery_zone_crawler	Ready		-
<input type="checkbox"/>	dl_landing_zone_crawler	Ready		Succeeded
<input type="checkbox"/>	dl_rocessing_zone_crawler	Ready		-

Crawler successfully starting

The following crawler is now starting: "dl_landing_zone_crawler"

AWS Glue > Crawlers > dl_landing_zone_crawler

dl_landing_zone_crawler

February 3, 2023

Crawler properties

Name dl_landing_zone_crawler	IAM role prefix-prod_Role_GlueCrawler	Database dl_landing_zone
Description -	Security configuration -	Lake Formation configuration -
Maximum table threshold -		

Advanced settings

Crawler runs Schedule Data sources Classifiers Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DP
February 3, 2023 at 01:15:31	February 3, 2023 at 01:16:26	55 s	Completed	

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings

Workgroup: primary

Data

Data source
AwsDataCatalog

Database
dl_landing_zone

Tables and views
Create

Filter tables and views

Tables (1)

customers	:
customer_id	string
customer_unique_id	string
customer_zip_code_prefix	string
customer_city	string
customer_state	string

Views (0)

Query 3 : X Query 4 : X

1 Select * from dl_landing_zone.customers

SQL Ln 1, Col 40

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 86 ms Run time: 997 ms Data scanned: 8.62 MB

Results (100+)

Copy Download results

2.1.2 Experiências vivenciadas

Ao longo da Sprint, algumas tarefas possuíram um nível de dificuldade menor de forma que a conclusão dessas foi relativamente rápida, quando comparadas às tarefas mais difíceis relativas a construção dos códigos de Terraform, responsáveis por disponibilizar os serviços necessários para arquitetura almejada. Existem alguns pontos a serem destacados:

- Quantidade de dados: Os dados coletados em sua maioria refletem a natureza dos dados de uma empresa de varejo e-commerce, mas apenas algumas tabelas devem ser utilizadas nesse projeto. Será interessante na próxima sprint verificar quais dessas tabelas de fato serão utilizadas.
- Qualidade de dados: Talvez valha a pena pensar se na próxima sprint pode ser válido acrescentar uma tarefa de limpeza de dados ou algo que possa garantir a qualidade dos dados em questão. Ou seja, na tarefa futura de “tratamento dos dados” talvez seja interessante dividir em formato e limpeza.

Sobre os códigos construídos em Terraform, vale salientar que muito do que foi construído, se deve em especial pela utilização do pacote de construção de infraestrutura IaC facilitado pelo Rony. A utilização do Github no processo de disparo e construção de serviços via “actions” (yaml) facilitou não só a construção dos serviços como a sua destruição e consequente acompanhamento de custos do projeto. Além disso, a detecção de erros foi muito bem entendida justamente por estar bem detalhada e disponibilizada no actions do github o que aumentou a produtividade.

Outro ponto a ser destacado é o fato de que por algumas vezes o serviço, por apresentar algum erro, teve que ser desligado manualmente, o que leva certo tempo.

2.2 Sprint 2

2.2.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.2.2 Experiências vivenciadas

2.3 Sprint 3

2.3.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.3.2 Experiências vivenciadas

3. Considerações Finais

3.1 Resultados

Por meio de um texto detalhado, apresente os principais resultados alcançados pelo seu Projeto Aplicado.

Cite os pontos positivos e negativos, as dificuldades enfrentadas e as experiências vivenciadas durante todo o processo.

3.2 Contribuições

Apresente quais foram as contribuições que o seu Projeto Aplicado trouxe para que o Desafio proposto fosse solucionado.

Cite, por exemplo, as inovações, as vantagens sobre os similares, as melhorias alcançadas, entre outros.

3.3 Próximos passos

Descreva quais são os próximos passos que poderão contribuir com o aprimoramento da solução apresentada pelo seu Projeto Aplicado.