UNIVERSIDADE FEDERAL DO MARANHÃO

Engenharia da Computação

Francicleia Rodrigues Cardoso

Implementação de Arquitetura Google Cloud para Análise e Visualização de Dados

São Luís, MA

2024

Francicleia Rodrigues Cardoso

Implementação de Arquitetura Google Cloud para Análise e Visualização de Dados

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Maranhão como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação.

Orientador: Prof. Dr.Haroldo Gomes Barroso Filho

São Luís, MA

2024

Francicleia Rodrigues Cardoso

Implementação de Arquitetura Google Cloud
para Análise e Visualização de Dados

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Maranhão como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação.

Trabalho _____ em São Luís, _____ de 2024:

_____

Prof. Dr.Haroldo Gomes Barroso Filho
Orientador

_____

Examinador

_____

Examinador

São Luís, MA
2024

Resumo

É crescente o número de lojas virtuais, em 2023 o comércio online registrou um faturamento de R$ 185.7 BI (ABCOMM, 2024). Com o número expressivo de comerciantes atuando no mundo virtual, a concorrência por clientes aumenta e a garantia da satisfação do cliente se torna um diferencial crucial para o sucesso da empresa. Seja para aprimorar o atendimento, criar campanhas de vendas mais personalizadas ou visualizar tendências ao longo do tempo, a tomada de decisão orientada a dados vem se mostrando cada vez mais uma vantagem competitiva significativa para as empresas. A presente pesquisa apresenta uma proposta de arquitetura de dados em nuvem utilizando a plataforma Google Cloud. Utilizando uma base de dados de pedidos realizados no ano de 2017 em marketplaces brasileiros, foi implementada a extração, tratamento e armazenamento e visualização de dados, além do emprego de inteligência artificial para sumarização de comentários de clientes para gerar insights sobre a experiência de compra e a satisfação.

O fluxo integrando a extração, transformação e carregamento (ETL) de dados foi implementado utilizando a linguagem *Python*, por meio de uma função no *Cloud Function*, e a execução automatizada foi programada através do *Cloud Scheduler*. Os dados foram armazenados no *BigQuery* e visualizados em um *dashboard* interativo criado no *Looker Studio*. A sumarização dos comentários foi realizada por meio de modelos de linguagem avançados, permitindo a identificação dos principais tópicos de *feedback* dos clientes. Os resultados mostraram que a arquitetura atendeu às necessidades propostas e é escalável, podendo ser aprimorada com ajustes para maior eficiência e confiabilidade.

Palavras-chave: Arquitetura em nuvem, Google Cloud, Análise de Dados, Visualização de dados

Abstract

The number of online stores is steadily increasing, with online commerce reaching a revenue of R$ 185.7 billion in 2023 (ABCOMM, 2024). As the number of merchants operating in the virtual world grows, competition for customers intensifies, and ensuring customer satisfaction becomes a crucial differentiator for a company's success. Whether to improve customer service, create more personalized sales campaigns, or visualize trends over time, data-driven decision-making has proven to be a significant competitive advantage for businesses. This research presents a proposal for a cloud-based data architecture using the Google Cloud platform. Using a dataset of orders placed in 2017 on Brazilian marketplaces, data extraction, processing, storage, and visualization were implemented, along with the use of artificial intelligence to summarize customer comments to generate insights into the shopping experience and satisfaction.

The workflow integrating data extraction, transformation, and loading (ETL) was implemented using Python, through a Cloud Function, and automated execution was scheduled using Cloud Scheduler. The data was stored in BigQuery and visualized in an interactive dashboard created in Looker Studio. The comment summarization was performed using advanced language models, allowing the identification of key customer feedback topics. The results showed that the architecture met the proposed needs and is scalable, with potential improvements for greater efficiency and reliability.

Keywords: Cloud architecture, Google Cloud, Data analysis, Data visualization

Lista de ilustrações

Sumário

1 Introdução

O comércio eletrônico tem promovido mudanças significativas no setor varejista, com as compras online se tornando cada vez mais comuns. Segundo a Associação Brasileira de Comércio Eletrônico (ABCOMM, 2024), em 2023 foi registrado um faturamento de R$ 185.7 BI em compras realizadas em plataformas de ecommerce e existiam cerca de 87.8 MI compradores online.

Assim como a demanda por produtos online, a oferta de fornecedores também tem crescido. Em um ambiente onde o consumidor tem tantas opções sem sair de casa, ganha quem alcançar um maior nível de satisfação e atender melhor às expectativas dos clientes. Segundo Wei et al. (2022), a qualidade do serviço é fundamental para manter o cliente engajado e garantir o uso contínuo dos produtos ou serviços de uma empresa ao longo do tempo.

Com relação à satisfação do cliente no ambiente online, Ilieva et al. (2022) afirmam que ela é determinada ao avaliar os seguintes aspectos: a experiência de compra no site, a interação com o suporte técnico, a entrega da mercadoria e, caso o e-commerce disponha, o uso do aplicativo móvel. Além disso, os autores destacam que a vantagem competitiva de uma empresa e o sucesso financeiro dos produtos e serviços oferecidos no meio digital são medidos pela satisfação do cliente.

Provost e Fawcett (2016), em seu livro Data Science para Negócios, afirmam que o primeiro passo para as empresas é estabelecer estruturas que permitam o armazenamento e processamento flexível de grandes volumes de dados. Somente após essa etapa, poderão identificar novas oportunidades ou aprimorar a eficiência de seus métodos de execução. Atualmente, um número crescente de empresas reconhece que parte do sucesso de seus negócios deriva das informações extraídas dos dados que processam. Graças à revolução da análise de big data, é possível obter uma compreensão cada vez mais profunda dos hábitos e preferências dos clientes (Wei et al., 2022).

Para extrair informações de grandes volumes de dados, diversas teorias e técnicas da área de inteligência artificial têm sido amplamente utilizadas, como o aprendizado de máquina, o aprendizado profundo e o processamento de linguagem natural. Taulli (2020) destaca que, por meio do uso de aprendizado de máquina, uma empresa pode utilizar os dados armazenados para identificar as abordagens mais eficazes e obter insights valiosos.

O presente trabalho se insere nesse contexto, com o objetivo de propor uma arquitetura em nuvem que possibilite a análise de dados e a extração de informações relacionadas a uma base de dados de vendas online no Brasil, extraída da plataforma Kaggle.

1.1 Objetivos

De maneira geral, o presente trabalho tem como objetivo propor uma arquitetura em nuvem que integre as etapas de extração, análise, armazenamento e visualização de dados, utilizando a plataforma Google Cloud. A proposta é aplicada aos dados de pedidos de marketplaces brasileiros do ano de 2017, extraídos da plataforma Kaggle. Para alcançar o objetivo geral, definimos os seguintes objetivos específicos:

a. Desenvolver uma arquitetura de armazenamento das bases de dados utilizadas no processo de análise e, posteriormente, a base de dados para consumo em dashboard em um ambiente cloud;
b. Realizar as etapas de tratamento e análise exploratória de dados;
c. Desenvolver um dashboard interativo para visualização de dados, possibilitando a geração de insights, centralização de indicadores e consultas interativas aos dados.

1.2 Organização do Trabalho

Este trabalho está dividido em seções, refletindo as etapas realizadas ao longo do desenvolvimento do mesmo. Após a introdução, será apresentada a fundamentação teórica que servirá como base para o entendimento das tecnologias e processos empregados. Em seguida, detalharemos a metodologia utilizada, discutiremos os resultados obtidos e, na conclusão, faremos uma reflexão sobre a execução do projeto e, principalmente, sobre a relevância da utilização de técnicas de classificação de sentimentos ao se trabalhar com dados de clientes.

2 Referencial Teórico

Contextualizando a discussão a seguir, a sessão começa com uma abordagem sobre os fundamentos da análise de dados. Em seguida, aborda-se a computação em nuvem e o Google Cloud, uma vez que esse será o ambiente em nuvem utilizado.

2.1 Análise de Dados

Os dados têm se tornado um aliado importante na tomada de decisão, seja ela operacional ou estratégica (Patil, 2011). Atualmente, a grande maioria dos modelos de negócio já possui uma cultura propícia à coleta de dados e muitos já são instrumentados para tal finalidade. A exploração de dados garante uma vantagem competitiva, possibilitando a previsão de cenários e a adoção de estratégias para impulsionar o negócio (Provost; Fawcett, 2016).

Para que os dados brutos se tornem úteis e possam ser utilizados para seus devidos propósitos, é necessário que eles passem por uma série de processos, que vão desde a coleta, tratamento, análise e, finalmente, a disponibilização. A integração de todos esses processos é conhecida como ciclo de vida do dado (Badia, 2021). Na primeira etapa, temos os dados brutos, que são aqueles que não foram submetidos a nenhuma forma de pré-processamento. É importante ressaltar que, nessa fase, ocorre a aquisição de dados, que podem estar em vários formatos e diferentes fontes.

A fase seguinte é o tratamento ou a preparação dos dados. Em geral, para que sejam aplicadas técnicas de análise, os dados precisam passar por algumas conversões. De acordo com Provost e Fawcett (2016), exemplos típicos de tratamento de dados incluem: conversão para formato tabular, tratamento de valores ausentes, conversão de dados, normalização de dados numéricos, entre outros.

Sobre o processo de análise, Amaral (2016) destaca que, “analisar dados é aplicar algum tipo de transformação nos dados em busca de conhecimento”. Ainda segundo o autor, a análise de dados pode ser dividida em duas categorias: análise explícita, quando o conhecimento e a informação estão disponíveis nos dados e a análise acontece apenas com o intuito de destacar esses dados; e análise implícita, em que, para que os dados recebam o destaque necessário e a informação possa ser extraída, é necessária a aplicação de técnicas como o aprendizado de máquina ou técnicas estatísticas. Finalizando o ciclo de vida do dado,

uma vez que ele tenha sido analisado, existem dois caminhos possíveis: o dado é armazenado em um sistema de armazenamento, pois pode ser útil futuramente; ou é deletado, caso não haja necessidade de mantê-lo.

2.2 Processamento de Linguagem Natural (PLN)

Vinculada à área de Inteligência Artificial e Linguística Computacional, o Processamento de Linguagem Natural (PLN) é o campo de pesquisa voltado para o desenvolvimento de métodos e sistemas capazes de realizar o processamento computacional da linguagem humana. Em outras palavras, no âmbito do PLN, o objetivo principal é a solução de problemas computacionais relacionados ao tratamento de um idioma, seja ele falado ou escrito (Caseli; Nunes, 2023).

Os seres humanos possuem a capacidade de gerar e entender linguagens, que na sua grande maioria, são não estruturadas e repletas de improvisações. No entanto, ao contrário de nós, os computadores são limitados ao entendimento de um conjunto de regras estruturadas. O PLN torna viável o processamento e análise desses dados não estruturados, possibilitando uma variedade de pesquisas e avanços na área de aprendizado de máquina. As possibilidades dentro do PLN variam desde tarefas como análise de sentimentos até geração de linguagem natural (Patel et al., 2020).

A análise de sentimentos é uma área dentro do PLN que possibilita a interpretação e classificação de emoções em geral, dividindo-as em três categorias: positivo, negativo e neutro. Com a utilização de algoritmos de análise de sentimentos, é possível prever, analisar e avaliar como os clientes se sentem sobre um determinado produto, por exemplo (Taherdost; Madanchian, 2023).

2.2.1 Modelos de Linguagem Grandes (LLM)

Modelos de Linguagem Grandes, do inglês *Large Language Model* (LLM), são modelos de inteligência artificial treinados em uma vasta quantidade de textos. Segundo Nascimento (2024), "esses modelos desenvolvem a habilidade de antecipar a próxima palavra em uma frase" e possuem aplicação em diversas áreas, como, por exemplo, a criação de textos que imitam a escrita humana ou até mesmo a geração de textos de forma autônoma.

Os LLMs possuem a habilidade de aprendizado em contexto, do termo em inglês *in-context learning* ou f*ew-shot learning*, o que significa que eles são capazes de executar tarefas para as quais não foram diretamente treinados. Tal habilidade torna possível a utilização desses algoritmos através de instruções em linguagem natural, os *prompts*, ou demonstrações de tarefas com exemplos, tudo isso fazendo uso do seu pré-treinamento, sem a necessidade de ajustes específicos (Taulli, 2020).

De acordo com Nascimento (2024), os LLMs "são frequentemente utilizados em análise de sentimentos devido à sua capacidade de processar e compreender grandes volumes de texto". Ainda segundo o autor, os modelos são capazes de identificar padrões linguísticos e contextuais que classificam os sentimentos em positivos, neutros ou negativos. A capacidade desses algoritmos de fazer interpretações de forma contextualizada e ampla em linguagem natural tem se mostrado eficaz em aplicações como "análise de sentimentos em redes sociais, avaliação de produtos, *feedback* de clientes, entre outros" (Nascimento, 2024).

2.3 Computação em Nuvem

A computação em nuvem, ou *cloud computing* em inglês, tem transformado a gestão e utilização de recursos computacionais ao oferecer flexibilidade, escalabilidade, segurança e opções de custo eficientes. Embora computadores mais potentes tenham se tornado mais acessíveis, o compartilhamento de resultados e o treinamento de um modelo com novos dados,por exemplo, continuam sendo mais difíceis quando esses dados estão armazenados em uma máquina local (Borra, 2024). Construída em uma arquitetura distribuída, a computação em nuvem distribui seus processos por vários servidores, em vez de concentrá-los em apenas um. Isso possibilita redundância de infraestrutura, garantindo que uma aplicação não sofra interrupções em sua execução (Ciaburro; Ayyadevara; Perrier, 2018).

Nos últimos anos, a demanda por serviços em nuvem tem aumentado significativamente. Empresas como Amazon, Microsoft, IBM e Google são algumas das que possuem suas próprias plataformas de computação em nuvem (Gupta; Mittal; Mufti, 2021).

Com relação à responsabilidade de gerenciamento, como administração de hardware, software, atualizações, segurança e monitoramento, existem dois segmentos principais no mercado. O primeiro é o de nuvem gerenciada, onde o provedor de computação em nuvem se responsabiliza pela gestão da infraestrutura e dos serviços, garantindo maior facilidade de uso, suporte e manutenção, com menor responsabilidade para o usuário. O segundo segmento

é o de nuvens não gerenciadas, onde a manutenção e configuração dos ambientes ficam a cargo do usuário. Esse modelo oferece maior flexibilidade e controle, permitindo uma personalização mais eficiente conforme as necessidades, mas também acarreta responsabilidades adicionais (Ciaburro; Ayyadevara; Perrier, 2018).

Além do tipo de gerenciamento, os ambientes em nuvem são categorizados em nuvens públicas, privadas ou híbridas. O primeiro modelo, nuvem pública, é caracterizado pelo compartilhamento de recursos entre clientes, e a localização exata da infraestrutura utilizada não é conhecida; apenas a região geográfica pode ser selecionada. Em nuvens públicas, o modelo de preço é chamado de "sob demanda", variando de acordo com o poder computacional utilizado, volume de dados armazenados e outras questões relacionadas à infraestrutura (Ciaburro; Ayyadevara; Perrier, 2018).

No caso de nuvens privadas, a arquitetura é projetada individualmente, podendo ser armazenada localmente ou em algum data center. A manutenção dessas nuvens é de total responsabilidade do cliente. Embora sejam mais caras por utilizarem hardware dedicado, oferecem maior controle sobre a infraestrutura. Já as nuvens híbridas combinam os dois modelos, aproveitando as vantagens de ambos (Ciaburro; Ayyadevara; Perrier, 2018).

2.3.1 Google Cloud Platform

A Google oferece a Google Cloud Platform (GCP), uma nuvem pública na qual os clientes compartilham os mesmos servidores. A GCP é uma nuvem não administrada, o que significa que o gerenciamento e a manutenção do ambiente ficam a cargo do usuário (Ciaburro; Ayyadevara; Perrier, 2018). Desde seu lançamento em 2018, a GCP tem expandido sua infraestrutura global e aumentado a diversidade de serviços oferecidos para atender às necessidades do mercado. Além de serviços de armazenamento, a plataforma oferece instâncias de máquinas virtuais, bancos de dados, serviços de rede, big data e análise de dados, aprendizado de máquina e inteligência artificial, entre outros (Borra, 2024).

Sobre o gerenciamento e acesso a recursos no GCP, Ciaburro, Ayyadevara e Perrier (2018), utiliza o termo "organização baseada em projetos". O usuário define um projeto, no qual serão agrupados os recursos e configurações necessários. O autor enfatiza que é inviável iniciar qualquer recurso sem especificar o projeto ao qual ele pertence. Cada projeto deve conter as seguintes configurações:

- Nome (definido pelo usuário);
- ID do projeto (sugerido pelo GCP, mas editável);
- Número do projeto (fornecido pelo GCP).

3 Metodologia

A presente seção é a descrição da abordagem metodológica empregada no desenvolvimento do projeto. A Figura 1 apresenta a arquitetura do projeto, bem como as tecnologias e ferramentas empregadas em cada etapa. Cada subseção a seguir é um detalhamento de como se deu o desenvolvimento de cada fase da arquitetura do projeto.

Figura 1 - Diagrama da Arquitetura Integrando os Serviços em Nuvem Utilizados

Fonte: Elaborado pela autora

3.1.1 Fontes de dados

As bases de dados utilizadas fazem parte do "Brazilian E-Commerce Public Dataset by Olist," um conjunto de informações públicas contendo dados de aproximadamente 100 mil pedidos de consumidores brasileiros no mercado de compras online, entre setembro de 2016 e setembro de 2018. Os dados originais foram extraídos no formato CSV (valores separados por vírgula) da plataforma Kaggle, "a maior comunidade de ciência de dados do mundo" (Kaggle, 2024), e armazenados em planilhas Google. Os bancos de dados presentes no dataset e que serão empregados ao longo do projeto estão listados a seguir:

- olist_customers_dataset.csv

- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv

Dentro da arquitetura, as fontes de dados são transformadas em planilhas *Google*, e as etapas seguintes foram desenvolvidas no GCP.

3.1.2 Consolidação das bases, análise de dados e armazenamento

O ponto central da arquitetura do projeto é o Cloud Function. Uma arquitetura serverless, ou seja, um ambiente de execução sem servidor que possibilita a criação e conexão de serviços na nuvem. Ativadas e desativadas conforme a necessidade de uso, as funções em nuvem têm o custo de uso calculado com base no tempo de execução do código, oferecendo ótimo custo-benefício e eficiência. Como não é necessário realizar ajustes na infraestrutura ou gerenciar servidores, o papel do usuário se limita a inserir o código e configurar o gatilho que dispara a execução da função (Google Cloud, 2023).

Como mencionado, não é necessária intervenção humana na execução da função; elas são ativadas por meio de gatilhos, que respondem a eventos na nuvem. Exemplos desses eventos incluem alterações em bancos de dados, criação de novos arquivos em um sistema de armazenamento ou acionamento da função pelo Cloud Scheduler (Google Cloud, 2023).

Para este projeto, a função foi criada utilizando a linguagem de programação *Python* e realiza a integração das bases de dados, leitura em um *dataframe*, transformação dos dados, criação de novas métricas e exportação da base resultante para uma tabela no *BigQuery*. A tarefa será executada de forma periódica, programada para ocorrer a cada três horas. Para isso, foi utilizado o *Cloud Scheduler*, um serviço gerenciado do *Google Cloud* que permite o agendamento de execução de tarefas em períodos ou horários específicos (Google Cloud, 2023).

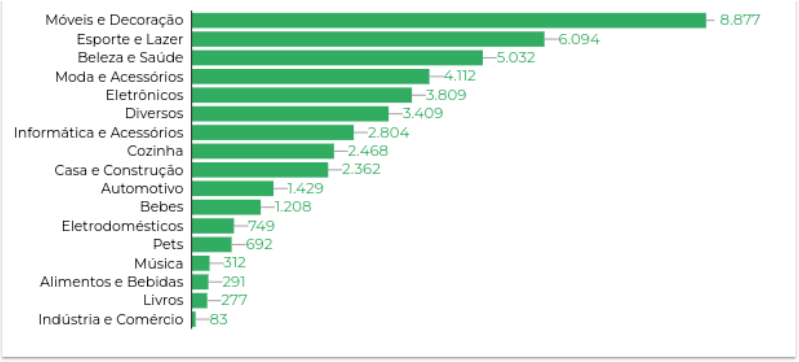
A primeira etapa da função corresponde a consolidação das bases de dados em um único *dataframe*. Logo em seguida deu-se início ao processo de preparação dos dados. Para o desenvolvimento deste projeto optou-se por trabalhar apenas com os dados de 2017, o que corresponde a um conjunto de dados com 46314 linhas e 31 colunas, com diversas

informações sobre os pedidos realizados, como status do pedido e preço; avaliação do pedido, incluindo nota, comentário e categoria do produto; dados dos clientes, como cidade e estado; e informações sobre os vendedores, também abrangendo cidade e estado.

Além da delimitação de período para análise de dados iniciais, optou-se por excluir as linhas de valores nulos da coluna product_category_name, sem o nome dos produtos na base de dados seria inviável realizar uma categorização adequada dos mesmos.

A coluna product_category_name contém um total de 73 categorias de produtos, em que muitas dessas categorias são de produtos com finalidades semelhantes e que podem ser agrupados. Foi realizado um agrupamento dessas categorias, reduzindo-as a 17 categorias principais. A Figura 2 ilustra como essas categorias foram organizadas e agrupadas.

Figura 2 - Distribuição de Pedidos por Categoria



Fonte: Elaborado pela autora

Com relação à localização geográfica tanto de clientes quanto de vendedores, a base inclui a cidade e o estado de ambos. No entanto, para entender a distribuição das compras por região, foram criadas as colunas regiao_cliente e regiao_estado, a partir do agrupamento dos estados brasileiros que compõem cada região.

Considerando que o cronograma logístico do pedido é um fator que afeta a satisfação do cliente, conforme discutido por Sutrisno, Andajani e Widjaja (2019), decidiu-se criar as seguintes métricas:

- total_dias_entrega_transportadora: o tempo gasto em dias entre a criação do pedido e a entrega para a transportadora;
- total_dias_entrega_cliente: o tempo gasto em dias entre a criação do pedido e a entrega para o cliente;

- total_dias_entrega_prevista: o total de dias entre a criação do pedido e a data de entrega prevista.

A base de dados apresenta comentários e avaliações de cada pedido realizado. Para sumarizar os comentários e analisar os principais tópicos trazidos, foi implementado uma função de sumarização agrupando os dados de pedidos por mês, status do pedido, região do cliente e nota dada na avaliação.

Na sumarização dos comentários, foram utilizados LLMs acessíveis por meio da *Vertex AI PaLM API*. A *Vertex AI* é uma plataforma do *Google* que permite o treinamento e a implantação de modelos de aprendizado de máquina e aplicativos de IA. Além disso, a plataforma possibilita a personalização de modelos LLMs para serem utilizados em aplicativos que incorporam tecnologia de IA (Google Cloud, 2024).

A implantação e a interação com modelos de linguagem de IA generativa por meio de um *Jupyter Notebook* são realizadas utilizando APIs. O *Google* disponibiliza as APIs *Vertex AI PaLM* e *Vertex AI Codey* para essa finalidade. Os modelos seguem uma nomenclatura padrão que indica o caso de uso e o tamanho do modelo. Por exemplo, o text-bison é um modelo de texto da família Bison. Bison é um tamanho de modelo PaLM que pode processar uma ampla variedade de tarefas de linguagem, incluindo classificação e resumo. Além do Bison, a família de modelos PaLM inclui o modelo de tamanho Unicorn, o maior da família, capaz de lidar com tarefas complexas como codificação, e o modelo Gecko, o menor, recomendado para a execução de tarefas simples (Google Cloud, 2024).

Para a utilização da *Vertex AI* foi realizada a conexão com a *Google AI Platform* na função, o *TextGenerationModel* foi instanciado, utilizando o modelo text-bison e a orientação para a sumarização dos principais tópicos foi dada por meio de uma instrução ao modelo. Após a criação da função, ela foi aplicada à coluna review_comment_message. Concluída a etapa de sumarização, a base de dados tratada foi armazenada.

A última etapa da função corresponde ao armazenamento da base resultante em uma tabela no *BigQuery* em um *dataset* específico para o projeto. Dentre os serviços de *big data* e análise de dados disponibilizados pelo GCP, o *BigQuery* é um sistema de armazenamento que opera no modelo "*serverless*". Segundo Borra (2024), o *BigQuery* é "otimizado para garantir consultas SQL rápidas em extensos conjuntos de dados, sendo ideal para análises em tempo real". Projetado para executar consultas ágeis com seleções e agregações em segundos, o *BigQuery* permite que empresas e desenvolvedores acessem e gerenciem grandes volumes de

dados sem se preocupar com a infraestrutura de *hardware* e *software* em tempo real (Ciaburro; Ayyadevara; Perrier, 2018).

O *BigQuery* é um serviço público acessível a todos por meio da nuvem que garante a proteção dos dados com múltiplos níveis de segurança, replicação em vários servidores e exportação simples e rápida (Ciaburro; Ayyadevara; Perrier, 2018). A ingestão de dados no *BigQuery* pode ser realizada em lotes, a partir de arquivos locais ou do *Cloud Storage*, ou ainda transmitidos em tempo real para dados gerados continuamente. A interação com o *BigQuery* pode ser feita de três maneiras: pela interface do console do *Google Cloud,* pela ferramenta de linha de comando do *BigQuery*, ou via API, acessível por bibliotecas cliente em linguagens como "*Python, Java, JavaScript* e *Go*, além da *API REST* e RPC para transformar e gerenciar dados" (Google, 2024).

Os dados armazenados em tabelas do BigQuery podem ser consumidos em diversas ferramentas, como *BI Engine, Looker Studio, Looker,* Planilhas *Google*, além de ferramentas de terceiros. No projeto em questão, os dados armazenados após a análise serão consumidos através de um *dashboard* na ferramenta *Looker Studio.*

3.1.3 Visualização de dados

Com o intuito de criar um painel de visualização de dados acessível a todos que precisarem consultá-lo e de tornar o projeto viável à medida que as bases de dados forem atualizadas com novos dados, foi desenvolvido um dashboard utilizando a ferramenta *Looker Studio.* Além de ser uma plataforma online, o que elimina a necessidade de instalação de software na máquina local, o *Looker Studio* oferece diversas possibilidades de conexão para integrar fontes de dados, o que inclui consultas personalizadas e tabelas do *BigQuery.*

Como mencionado anteriormente, essa base de dados contém informações de clientes, pedidos e vendedores, permitindo a realização de diversas análises. Neste trabalho, optou-se por focar no comportamento de compra e no nível de satisfação dos clientes em relação aos pedidos realizados. Para tal as visualizações desenvolvidas foram:

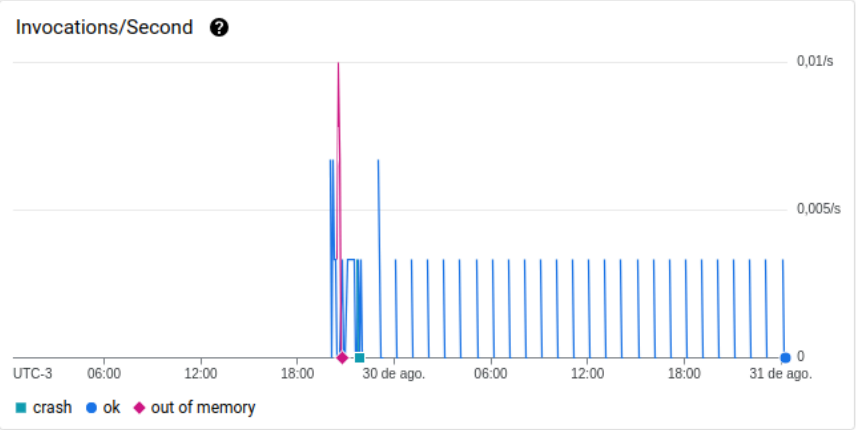
- Distribuição de pedidos realizados por status do pedido
- Total de Pedidos Realizados por Mês ao Longo do Tempo
- Distribuição de Pedidos por Categoria
- Distribuição do % de Avaliações por Pedidos Realizados
- Relação entre Avaliações e Status do Pedido

- Relação entre Avaliações e Comentários
- Mediana do Tempo de Entrega x Expectativa
- Relação entre Total de Pedidos e Valor Pago
- Sumarização dos comentários

4. Resultados

O presente estudo explorou a implementação de uma arquitetura em nuvem integrando a extração, manipulação de dados, armazenamento e visualização. O ponto principal da arquitetura proposta é a execução de uma *Cloud Function* onde todas as etapas citadas são gerenciadas e os serviços em nuvem necessários são invocados. Como citado anteriormente, o gatilho para a execução da função foi programado para ser executado a cada três horas. Como pode ser observado na Figura 3 a seguir, após a implantação a execução ocorreu de forma esperada, sem erros ou extrapolação da memória alocada.

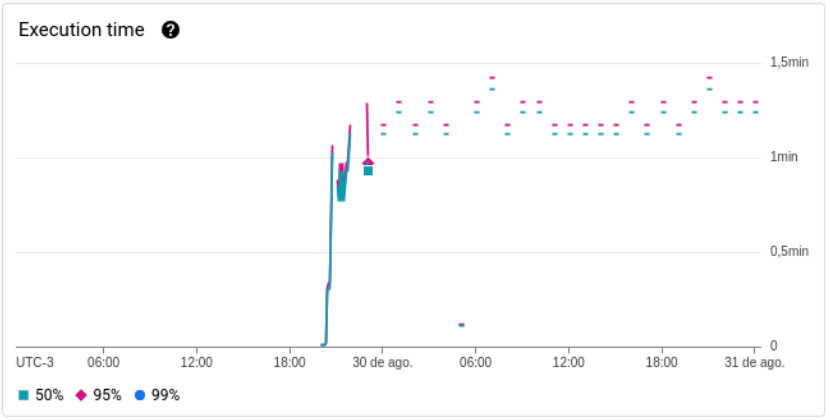
Figura 3 - Execução Automatizada da Função com o Google Scheduler



Fonte: Elaborado pela autora

O custo da utilização do Cloud Function está relacionado com o tempo de execução da função. Como pode ser observado na Figura 4, em relação ao tempo de execução, o maior tempo não ultrapassa a marca de 1,5 minutos.

Figura 4 - Tempo de Execução da Função no Cloud Functions



Fonte: Elaborado pela autora

Os dados tratados foram armazenados no *BigQuery*, e o painel de visualização foi desenvolvido com um total de nove visualizações. A primeira visualização, Figura 5, é a distribuição do total de pedidos por status, levando em consideração o total de pedidos realizados, a visualização apresenta o total de pedidos em cada status.

Figura 5 - Total de Pedidos por Status

Fonte: Elaborada pelo autora

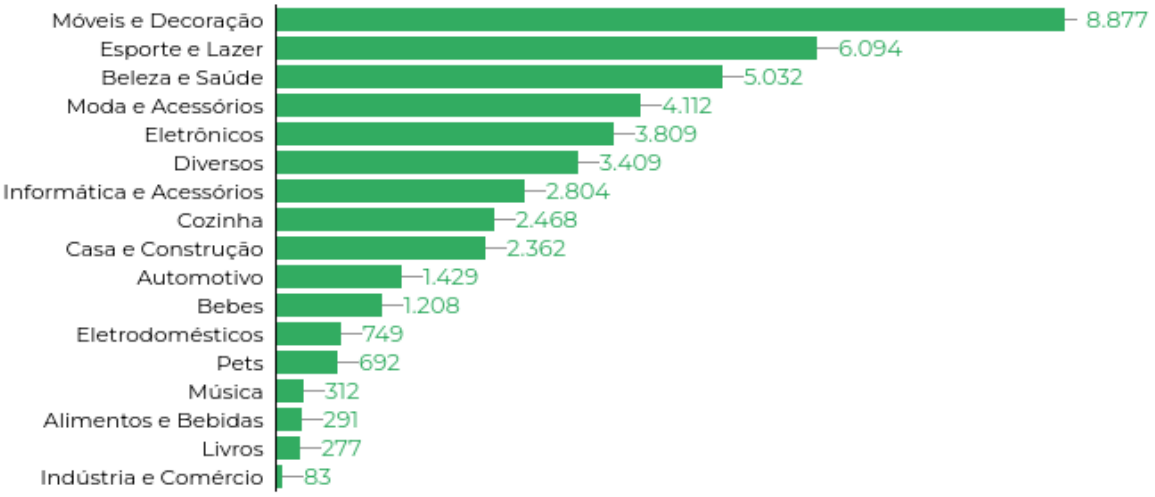
A Figura 6 apresenta a segunda visualização que corresponde ao total de pedidos realizados ao longo do ano de 2017.

Figura 6 - Total de Pedidos Realizados por Mês ao Longo do Tempo

Fonte: Elaborado pela autora

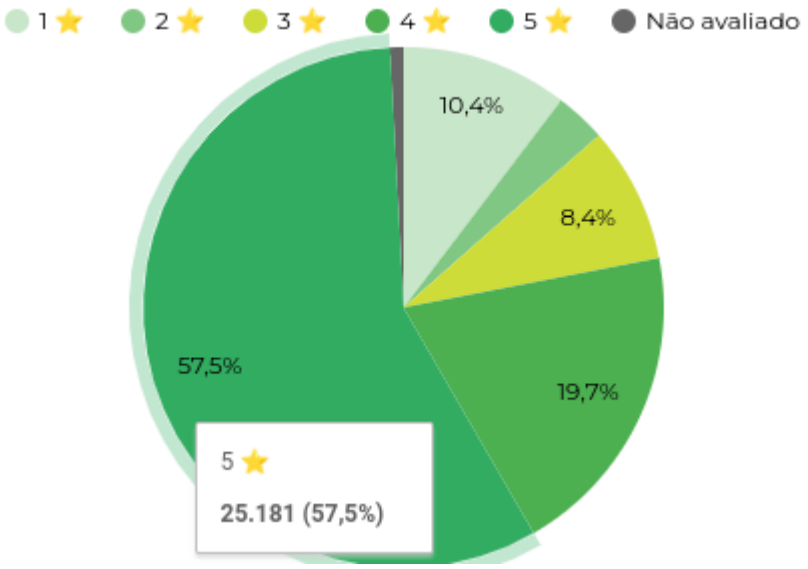
A Figura 7 apresenta o total de pedidos por categoria, enquanto a Figura 8 representa o percentual de avaliações realizadas. As notas de avaliação variam de 1 a 5, sendo que também existe a possibilidade de um pedido não receber avaliação. Observa-se que 57,5% dos pedidos foram avaliados com nota máxima, e 0,8% dos pedidos não receberam qualquer avaliação.

Figura 7 - Distribuição de Pedidos por Categoria



Fonte: Elaborado pela autora

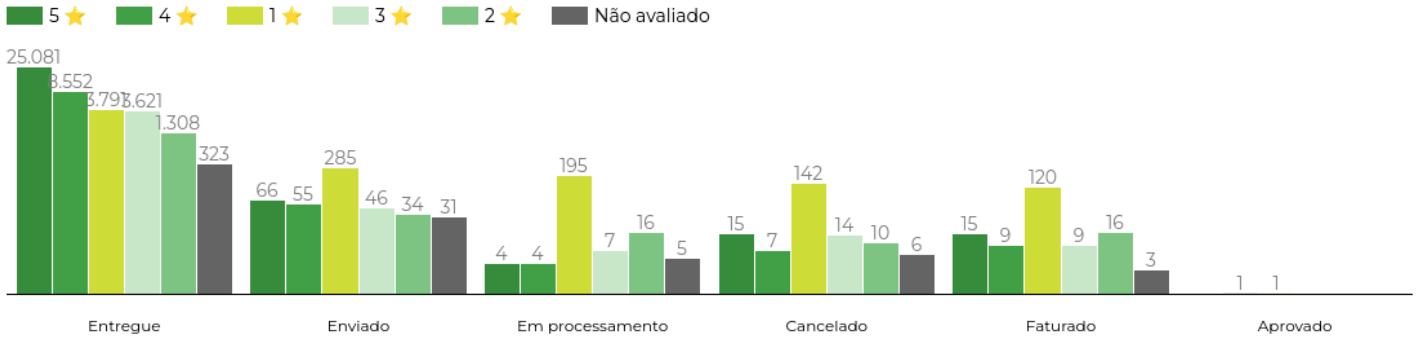
Figura 8 - Distribuição do % de Avaliações por Pedidos Realizados



Fonte: Elaborado pela autora

A Figura 9 mostra a relação entre os status do pedido e as notas atribuídas em cada avaliação. Observa-se que, com exceção do status "entregue", todos os demais status apresentam um número maior de notas mínimas em comparação com as outras notas.

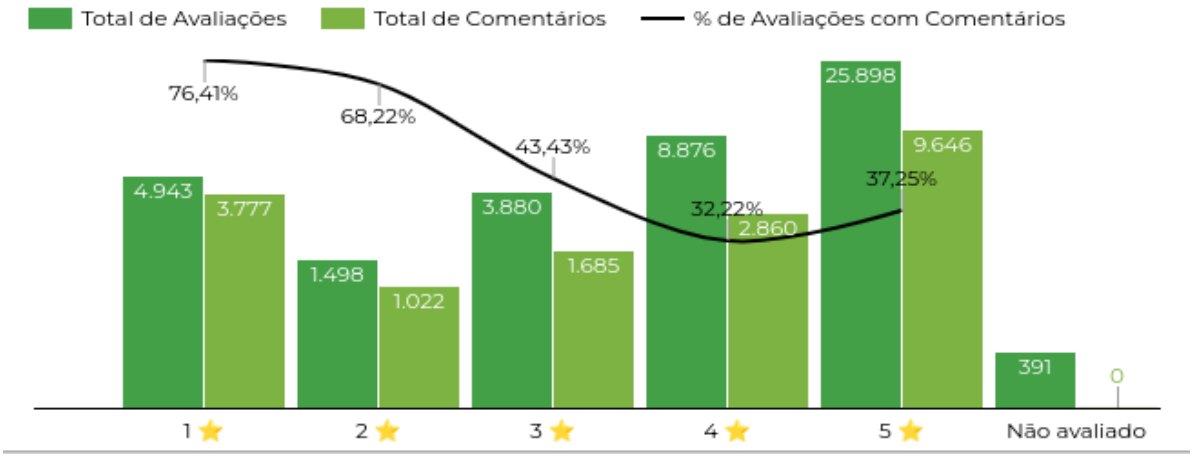
Figura 9 - Relação entre Avaliações e Status do Pedido



Fonte: Elaborado pela autora

A Figura 10 apresenta a relação entre o total de pedidos avaliados e o total de pedidos que receberam um comentário, distribuídos por nota. Observa-se que, embora a nota máxima tenha recebido o maior número de avaliações, a nota mínima possui o maior percentual de pedidos avaliados e comentados: 76,41% dos pedidos avaliados com nota 1 receberam um comentário.

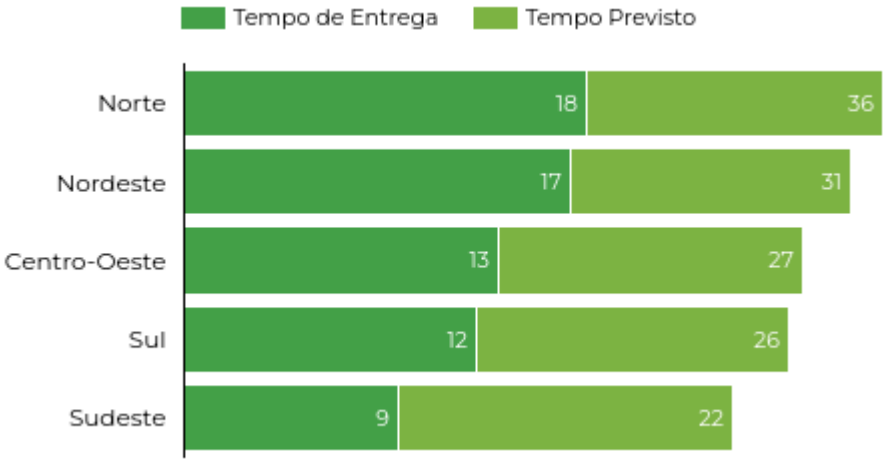
Figura 10 - Relação entre Avaliações e Comentários



Fonte: Elaborado pela autora

Na Figura 11, são apresentados o tempo mediano de entrega do pedido e o tempo estimado de entrega. Como a maioria dos pedidos é oriunda da região Sudeste, as médias para as regiões Norte e Nordeste são superiores em comparação às outras regiões.

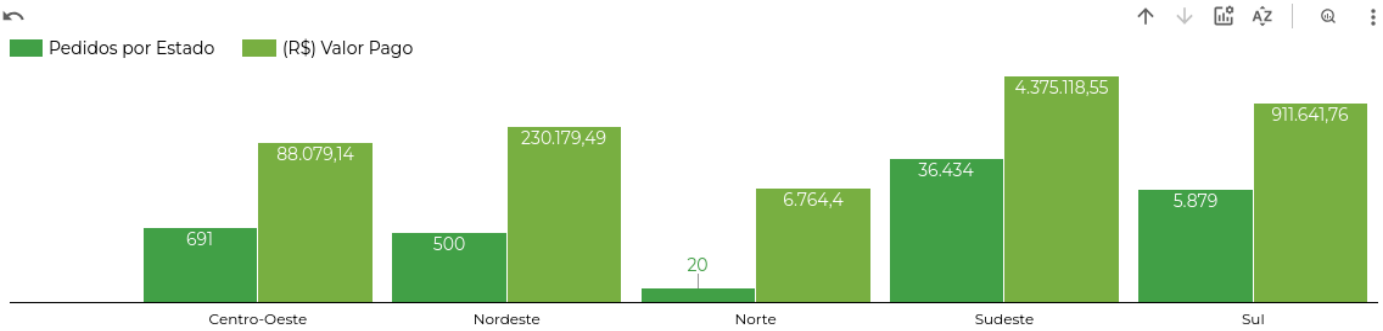
Figura 11 - Mediana do Tempo de Entrega x Expectativa



Fonte: Elaborado pela autora

A Figura 12 apresenta a relação entre o número de pedidos realizados por região ou estado e o valor pago. As regiões Sul e Sudeste destacam-se tanto pelo maior número de pedidos recebidos quanto pelo maior valor arrecadado.

Figura 12 - Relação entre Total de Pedidos e Valor Pago por Estado/Região



Fonte: Elaborado pela autora

A sumarização dos comentários de cada avaliação, Figura 13, foi realizada utilizando o LLM text-bison. O prompt fornecido ao modelo solicitava a listagem dos cinco principais tópicos, agregados por mês e ano, região do cliente e avaliação do pedido. O resultado da sumarização dos comentários permitiu identificar os principais critérios relacionados às notas atribuídas nas avaliações, além de destacar os pontos-chave da experiência de compra relatada pelos clientes, por região, período do ano e avaliações. A análise dos comentários é

fundamental, pois permite que as empresas identifiquem os principais fatores que influenciam a satisfação do cliente e desenvolvam ações estratégicas e personalizadas, cada vez mais alinhadas com as expectativas e necessidades dos mesmos.

Figura 13 - Sumarização dos Comentários

Data	Região do Cliente	Avaliação	5 Tópicos Principais
			3. Recomendação para colocar caixa frágil. 4. Aconteceu com o cliente. 5. Foto.
2017-08	Norte	3.0	1. Entrega rápida 2. Produto chegou sem problemas 3. Recomendação da loja 4. Satisfação do cliente 5. Experiência positiva
2017-08	Norte	4.0	**Principais tópicos das avaliações dos clientes:** 1. Produto com defeito. 2. Dificuldade de montagem. 3. Demora na entrega. 4. Produto de boa qualidade. 5. Entrega dentro do prazo.
2017-08	Norte	5.0	**1. Entrega dentro do prazo** **2. Atendimento excelente** **3. Produto de qualidade** **4. Preço justo** **5. Entrega rápida**

Fonte: Elaborado pela autora

O painel de visualização possui filtros interativos que permitem uma personalização dos dados de acordo com os objetivos do usuário, possibilitando uma análise mais direcionada.

Figura 14 - Filtros para seleção no painel

Fonte: Elaborado pela autora

5. Conclusão

A tomada de decisão baseada em dados concretos e variáveis históricas do negócio gera uma vantagem competitiva significativa para uma empresa, e é crescente a necessidade da extração e análise de dados gerados ao longo do tempo. Uma vez que a volumetria desses dados continua a crescer, torna-se essencial que todos os fatores para adequar uma cultura organizacional a esse cenário sejam considerados da melhor forma possível.

O objetivo deste artigo foi propor uma arquitetura de serviço em nuvem que realizasse as etapas de extração, transformação e carregamento de dados, integrando serviços de forma que a entrada e o tratamento de novos dados pudessem ser realizados de forma automatizada; que fosse possível a visualização dos dados em um dashboard e a sumarização de comentários através de LLMs. A arquitetura foi projetada fazendo uso de serviços da plataforma Google Cloud. A execução automatizada foi possível com o agendamento de eventos utilizando o Cloud Scheduler, que, por sua vez, habilita um gatilho para a execução de uma Cloud Function. A invocação dos outros serviços é realizada na Cloud Function e, com o término da sua execução, os dados são armazenados no BigQuery e consultados, por meio de visualizações gráficas, em um dashboard no Looker Studio.

É possível realizar a implementação de melhorias com o objetivo de tornar sua execução mais eficiente e otimizada, gastando o mínimo de recursos possíveis. Como sugestão a essas melhorias, temos: aumento do intervalo de execução e ajuste na lógica do código para extrair das bases apenas o que ainda não foi tratado. Com relação à confiabilidade dos dados, em um ambiente de produção, o ideal seria substituir as planilhas por ferramentas de armazenamento de dados mais confiáveis. Uma sugestão seria o Google Cloud Storage, que forneceria um controle de acesso granular aos dados.

De maneira geral, a arquitetura atende às necessidades propostas. Uma vez que o modelo em questão faz uso de serviços de nuvem, o modelo de arquitetura proposto é auto escalável e pode lidar com grandes volumes de dados e, com alguns ajustes, pode ser adaptado e utilizado em diversos cenários em que a tomada de decisão orientada a dados seja necessária.

Referências

ILIEVA, Galina et al. Customer Satisfaction in e-Commerce during the COVID-19 Pandemic. Systems, v. 10, n. 6, p. 213, 2022.

WEI, Li et al. Big data-driven personalization in e-commerce: algorithms, privacy concerns, and consumer behavior implications. *International Journal of Applied Machine Learning and Computational Intelligence*, v. 12, n. 4, p. 25, 2022.

GOOGLE. BigQuery: visão geral da consulta. Disponível em: https://cloud.google.com/bigquery/docs/query-overview?hl=pt-br. Acesso em: 16 ago. 2024.

REMESSA ONLINE. Google Data Studio: o que é e como utilizar? Disponível em: https://www.remessaonline.com.br/blog/google-data-studio/. Acesso em: 16 ago. 2024.

ASSOCIAÇÃO BRASILEIRA DE COMÉRCIO ELETRÔNICO (ABCOMM). *Números do E-commerce Brasileiro*. Dados ABComm, 2024. Disponível em: https://dados.abcomm.org/numeros-do-ecommerce-brasileiro. Acesso em: 31 ago. 2024.

SICHMAN, Jaime Simão. Inteligência artificial e sociedade: avanços e riscos. Estudos Avançados, São Paulo, v. 35, n. 101, p. 37-50, 2021.

IBM. Modelos de linguagem de grande escala. Disponível em: https://www.ibm.com/topics/large-language-models. Acesso em: 26 jun. 2024.

GOOGLE CLOUD. Introdução à Vertex AI: plataforma unificada. Disponível em: https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform?hl=pt-br. Acesso em: 26 jun. 2024.

GOOGLE CLOUD. Vertex AI: visão geral dos modelos de linguagem. Disponível em: https://cloud.google.com/vertex-ai/generative-ai/docs/language-model-overview?hl=pt-br. Acesso em: 26 jun. 2024.

CIABURRO, Giuseppe; AYYADEVARA, V. Kishore; PERRIER, Alexis. *Hands-on machine learning on Google Cloud Platform: Implementing smart and efficient analytics using Cloud ML Engine*. Packt Publishing Ltd, 2018.

BADIA, Antonio. *SQL for Data Science*. Springer, 2021.

AMARAL, Fernando. *Introdução à ciência de dados: mineração de dados e big data*. Rio de Janeiro: Alta Books Editora, 2016.

PROVOST, Foster; FAWCETT, Tom. *Data Science for Business*. Rio de Janeiro: Alta Books, 2016.

TAULLI, Tom. *Introdução à inteligência artificial: uma abordagem não técnica*. São Paulo: Novatec, 2020.

SUTRISNO, Angeline; ANDAJANI, Erna; WIDJAJA, Fitri Novika. *The effects of service quality on customer satisfaction and loyalty in a logistics company. KnE Social Sciences*, p. 85–92, 2019.

GOOGLE CLOUD. Google Cloud Functions: visão geral. Google Cloud, 2023. Disponível em: https://cloud.google.com/functions/docs/concepts/overview?hl=pt-br. Acesso em: 20 ago. 2024.

GOOGLE CLOUD. Google Cloud Functions: visão geral. Google Cloud, 2023. Disponível em: https://cloud.google.com/functions/docs/concepts/overview?hl=pt-br. Acesso em: 20 ago. 2024.

GOOGLE CLOUD. Cloud Scheduler. Google Cloud, 2023. Disponível em: https://cloud.google.com/scheduler?hl=pt-br. Acesso em: 20 ago. 2024.

KASTURIA, Vishesh; SHARMA, Shanu; SHARMA, Sachin. *Automatic product saleability prediction using sentiment analysis on user reviews.* In: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2020. p. 102-106.

PATIL, Dhanurjay J. *Building data science teams.* O'Reilly Media, Inc., 2011.

BORRA, Praveen. *A Survey of Google Cloud Platform (GCP): Features, Services, and Applications.* International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), v. 4, p. 191-199, 2024.

GUPTA, Bulbul; MITTAL, Pooja; MUFTI, Tabish. *A review on Amazon Web Service (AWS), Microsoft Azure & Google Cloud Platform (GCP) services.* In: Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD 2020), 27-28 fevereiro 2020, Jamia Hamdard, New Delhi, India. 2021.

SANTOS, Fernando Chagas; CARVALHO, Cedric Luiz de. *Aplicação da inteligência artificial em sistemas de gerenciamento de conteúdo.* Instituto de Informática, Universidade Federal de Goiás, 2008.

PATEL, Mihir et al. *Exploratory Data Analysis and Sentiment Analysis on Brazilian E-Commerce Website.* 2020. Tese de Doutorado – University of New York Polytechnic Institute, Utica/Rome, Utica, New York.

TAHERDOOST, Hamed; MADANCHIAN, Mitra. *Artificial intelligence and sentiment analysis: A review in competitive research.* Computers, v. 12, n. 2, p. 37, 2023.

CASELI, Helena M.; NUNES, Maria das Graças V.; PAGANO, Adriano. *O que é PLN.* In: CASELI, H. M.; NUNES, M. G. V. (org.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 28 jun. 2024.

NASCIMENTO, Jefferson Rodrigues do. *Exploração de técnicas de engenharia de prompt para aprimorar os resultados do uso de LLM no TCMRio.* 2024. Trabalho de Conclusão de Curso – Universidade Federal do Rio Grande do Norte, Natal.

KAGGLE. Brazilian E-Commerce Dataset. Disponível em: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Acesso em: 12 jun. 2024.