FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA - UNIFOR
CENTRO DE TECNOLOGIA E ENGENHARIA
CURSO MBA EM ENGENHARIA DE DADOS

TRABALHO FINAL PARA DISCIPLINA: ENGENHARIA DE DADOS EM NUVEM

FELIPE ALVES DA SILVA

Fortaleza-CE
2024

SUMÁRIO

1 INTRODUÇÃO

O setor de varejo tem experimentado uma transformação significativa nos últimos anos, impulsionada pela crescente utilização de dados para otimizar processos críticos, como a previsão de demandas e a gestão de estoques. A integração de análises avançadas de dados no varejo tem como objetivo melhorar a eficiência operacional, personalizar a experiência do cliente e reduzir custos operacionais. No entanto, mesmo com diversos avanços tecnológicos, diversos desafios relação à precisão das previsões de demanda ainda estão presentes, e frequentemente resultam em problemas como excessos ou deficits em estoque. Esses erros de previsão não apenas afetam a rentabilidade, mas também prejudicam a experiência do cliente, tornando-se um real problema em relação à eficiência do setor.

Neste contexto, o trabalho propõe o projeto de um sistema de recomendação baseado em dados, com a finalidade de aprimorar a gestão de estoques no varejo. A solução proposta faz uso de uma arquitetura em nuvem escalável e segura, capaz de integrar diferentes fontes de dados, incluindo históricos de vendas, comportamentos de consumo e comportamento de mercado. A principal intenção é garantir que os produtos estejam disponíveis nas quantidades ideais, minimizando tanto os custos operacionais quanto o impacto negativo nas operações e no atendimento ao cliente. Com isso, busca-se criar uma solução que não apenas otimize o inventário, mas também proporcione uma experiência mais satisfatória e eficiente para o consumidor final.

2 DESENVOLVIMENTO

2.1 DEFINIÇÃO DO PROBLEMA

Como já percebido com as informações anteriores, o setor selecionado foi o varejo devido ao grande cenário de atuação e o entendimento de crescente necessidade de analises de cenários.
A má previsão de demanda pode resultar em dois cenários problemáticos: o excesso de estoque, que gera custos elevados com armazenamento; e a escassez de produtos, que compromete as vendas e afeta negativamente a satisfação do cliente. Esses problemas aumentam devido às limitações das infraestruturas tradicionais de T.I, que muitas vezes não são capazes de processar grandes volumes de dados em tempo real ou de realizar análises preditivas de maneira ágil. Além disso, muitos sistemas de gestão de estoques ainda dependem de processos manuais e não conseguem integrar dados provenientes de diversas fontes, o que dificulta a adaptação rápida a flutuações no mercado.
Diante desse cenário, o objetivo deste trabalho é desenvolver uma solução inovadora baseada em tecnologias da AWS, com foco na integração e análise contínua e em tempo real dos dados. Ao utilizar ferramentas como Amazon Kinesis, AWS Glue, Amazon Redshift e Amazon SageMaker, a proposta é melhorar a precisão das previsões de demanda, otimizar os níveis de estoque e reduzir os custos operacionais. A adoção de uma abordagem em nuvem visa superar as limitações das infraestruturas tradicionais, proporcionando maior escalabilidade, agilidade e eficiência no processo dentro do ambiente de varejo.

2.2 PROCESSO DE COLETA, PROCESSAMENTO E ARMAZENAMENTO DE DADOS

A coleta de dados será realizada de maneira estruturada e contínua para garantir que a base de informações utilizada pelo sistema de recomendação seja precisa, atualizada e abrangente. O processo de coleta envolverá a integração de múltiplas fontes de dados, cada uma contribuindo com informações críticas para a previsão de demanda e otimização do estoque. O fluxo de coleta de dados será dividido nas seguintes etapas:

1. Identificação e Conexão com Fontes de Dados

Como primeiro passo, é necessário identificar as fontes de dados relevantes para o processo de coleta. Podemos seguir na analise da seguinte estrutura de dados:

- **Dados históricos de vendas**: Dados extraídos dos sistemas de PDV (pontos de venda) e plataformas de e-commerce, onde informações vitais estão presentes como produtos vendidos, quantidade, data, localização e etc.

- **Tendências de mercado**: Informações obtidas através de *web scraping* de fontes externas, como sites de análise de mercado, redes sociais e APIs de fornecedores de dados que fornecem informações sobre mudanças sazonais, lançamentos de produtos, campanhas promocionais e até mesmo comportamento do consumidor.

- **Dados logísticos e de fornecedores**: Informações obtidas através dos sistemas locais ou até mesmo diretamente de APIs dos fornecedores ou sistemas de ERP que oferecem detalhes sobre prazos de entrega, disponibilidade de estoque e custos.

2. Captura e Transmissão de Dados em Tempo Real

A coleta de dados será otimizada para processar informações em tempo real, utilizando ferramentas como **Apache Kafka** ou **Amazon Kinesis**, as quais permitirão capturar dados instantaneamente das fontes de dados citados anteriormente, coletando dados de estoque e vendas em tempo real. Esse fluxo contínuo de dados possibilitará atualizações rápidas das previsões de demanda, ajustando o modelo conforme novos dados entram no sistema.

3. Integração e Armazenamento de Dados

A integração de dados vindos de diferentes fontes será realizada por meio de processos de ETL utilizando ferramentas nativas da plataforma AWS, como o **AWS Glue**. A fase de extração abrangerá tanto a captura de dados em tempo real quanto em micro-batch, conforme a necessidade do fluxo de trabalho. Durante a transformação, serão aplicadas técnicas de

limpeza e preparação dos dados (Exemplos: remoção de duplicações, preenchimento de valores ausentes, normalização de formatos e etc, além de agregações dos dados por diferentes categorias, datas e regiões, percepções de outliers, etc). Após essa etapa, os dados limpos e transformados serão carregados em sistemas de armazenamento da AWS, como o **Amazon S3**, para dados não estruturados, e o **Amazon Redshift**, para dados estruturados, permitindo a realização de análises detalhadas e a aplicação de modelos preditivos.

4. Armazenamento e Leitura de Dados

Seguindo a pespectiva de padrão para dados não estruturados sendo destinado ao **Amazon S3** (configurado como um **Data Lake)** e para os dados estruturados sendo destinados ao **Amazon Redshift (**configurado como **Data Warehouse**), para dados semi-estruturados e de alta disponibilidade, como as interações dos clientes em plataformas de e-commerce, será empregado o **Amazon DynamoDB**, um banco de dados **NoSQL** que se destaca pela baixa latência e alta performance em consultas rápidas, atendendo de forma eficaz às exigências de acesso em tempo real.

5. Validação e Monitoramento de Qualidade de Dados

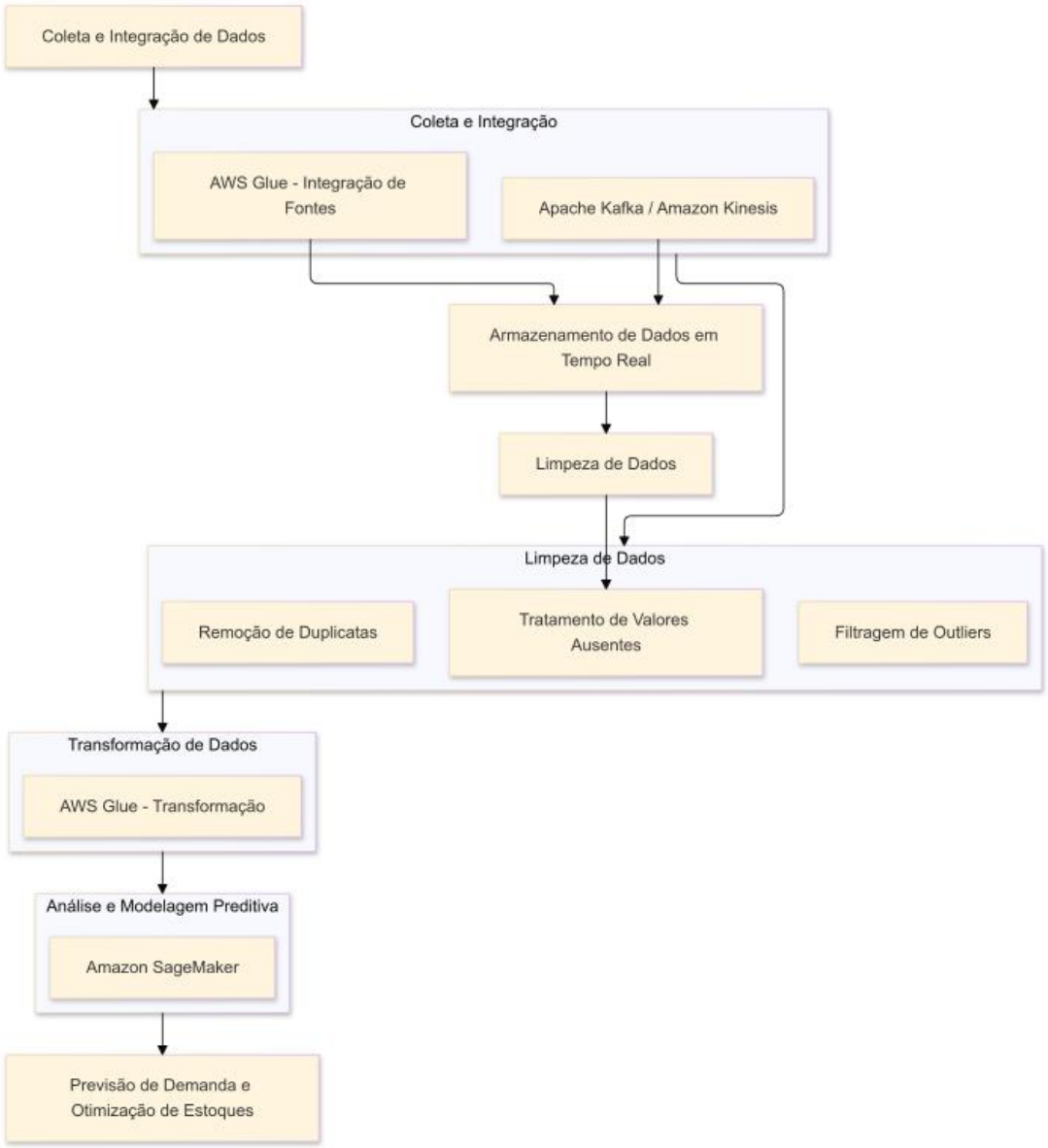
Ao longo de todo o processo de coleta e integração de dados, será estabelecido um processo contínuo de validação e monitoramento da qualidade, utilizando as ferramentas nativas da AWS. O **Amazon CloudWatch** será fundamental para assegurar que os dados coletados atendam aos critérios de consistência, integridade e alinhamento com os padrões de qualidade previamente definidos para o projeto. Esse monitoramente incluirá a identificação em tempo real de anomalias, como variações inesperadas nas vendas ou na disponibilidade de estoque, que possam indicar falhas na coleta de dados ou erros sistêmicos. Além disso, alertas automáticos serão configurados no CloudWatch de modo que seja notificado prontamente aos responsáveis sempre que qualquer desvio nos padrões de qualidade for identificado, permitindo uma resposta rápida e a correção de eventuais problemas.

6. Segurança e Conformidade com Regulamentações

Durante a coleta de dados, a proteção das informações sensíveis será garantida em conformidade com as regulamentações de privacidade, como a LGPD e o GDPR. A criptografia de dados será implementada com o **AWS KMS**, e o **AWS CloudTrail** será utilizado para auditoria e rastreabilidade de acessos, assegurando conformidade e segurança contínua.

2.3 DIAGRAMA DE PROCESSAMENTO DE DADOS

O processamento de dados será realizado por meio de um pipeline eficiente e escalável na AWS. No diagrama a seguir, resumo o processo em quatro etapas principais:



1. **Coleta e Integração de Dados**: Dados são coletados de várias fontes em tempo real e integrados por meio de ferramentas como Apache Kafka e serviços de ETL.
2. **Limpeza de Dados**: A qualidade dos dados é garantida, removendo duplicatas, tratando valores ausentes e filtrando outliers.
3. **Transformação de Dados**: Dados de diferentes fontes são normalizados e agregados para facilitar a análise.
4. **Análise e Modelagem Preditiva**: Modelos de machine learning são aplicados para prever a demanda e otimizar a gestão de estoques.

2.4 CONSIDERAÇÕES DE SEGURANÇA E GOVERNANÇA DE DADOS

A segurança e governança dos dados são aspectos fundamentais, e neste projeto, especialmente devido à natureza sensível das informações envolvidas, é preciso ter diversos tratamentos rigorosos pois há diversos dados sensíveis, como dados de clientes e transações comerciais. Com o intuito de garantir a proteção e a conformidade com as regulamentações de privacidade, como a LGPD e o GDPR, serão adotadas práticas rigorosas de segurança de dados na infraestrutura AWS:

1. **Proteção de Dados Pessoais:** A criptografia dos dados, tanto em momento de trânsito quanto em armazenamento, será implementada utilizando protocolos seguros, com o AWS KMS (Key Management Service) para gerenciar as chaves de criptografia. Isso assegura que os dados sensíveis sejam protegidos em todas as fases de processamento e armazenamento.

2. **Anonimização e Pseudonimização:** Para reduzir os riscos de exposição, os dados sensíveis serão anonimizados ou pseudonimizados por meio de ferramentas da AWS, como o AWS Glue DataBrew. Essas técnicas ajudam a proteger a identidade dos indivíduos, mantendo a utilidade dos dados para análises.

3. **Controle de Acesso e Autorização:** O controle de acesso será rigorosamente gerido através do AWS IAM (Identity and Access Management), permitindo a definição de permissões granulares. Isso garante que apenas usuários e serviços autorizados possam acessar dados sensíveis, minimizando o risco de acessos não autorizados.

4. **Políticas de Governança de Dados:** Políticas claras de governança de dados serão estabelecidas utilizando o AWS Lake Formation, ferramenta que facilita o gerenciamento e controle de dados em Data Lakes. Essa abordagem assegura a integridade, qualidade e rastreabilidade dos dados ao longo de seu ciclo de vida.

5. **Auditoria e Monitoramento:** Para garantir a conformidade contínua com as regulamentações, o AWS CloudTrail e o Amazon CloudWatch serão empregados para monitorar e auditar o uso dos dados. Essas ferramentas permitem rastrear acessos, manipulações e identificar comportamentos suspeitos, proporcionando uma camada adicional de segurança e controle.

Todas estas práticas associadas a uma boa poítica de segurança de dados implementada pela própria empresa visam assegurar que os dados sejam tratados de forma ética, segura e em plena conformidade com as normas legais, protegendo não apenas a privacidade dos consumidores, mas também a integridade do sistema de processamento de dados como um todo.

3 CONCLUSÃO

Este projeto teve como objetivo elaborar um sistema de recomendação com foco para otimizar a gestão de estoques no varejo, tendo como diretriz principal o uso de tecnologias da AWS. A solução envolve integração de dados de vendas, comportamento do consumidor e tendências de mercado para prever a demanda de produtos, minimizando excessos e déficits de estoque. Ferramentas como Amazon Kinesis, AWS Glue, Amazon Redshift e Amazon SageMaker foram essenciais para criar um sistema escalável e eficiente.

Foi priorizado uma pespectiva de obter melhorias na precisão das previsões de demanda, otimização dos níveis de estoque e redução de custos com armazenamento, além de ter como objetivo final também aprimorar a experiência do cliente. A segurança e a governança dos dados foram asseguradas, de forma ter o foco em agir com conformidade com as regulamentações de privacidade de dados.

Apesar dos ganhos percebidos, é possível identificar desafios como a necessidade de atualizar continuamente o modelo preditivo e a detecção de anomalias nos dados. Para isto, a implementação de monitoramento com Amazon CloudWatch garante a adaptação do sistema às mudanças nas condições do mercado.

O projeto oferece uma base sólida para o aprimoramento contínuo da gestão de estoques, com planos de refinamento do modelo e expansão das fontes de dados. A infraestrutura da AWS se demonstra ser uma plataforma eficaz para lidar com os desafios do gerenciamento de grandes volumes de dados.

4 REFERÊNCIAS BIBLIOGRÁFICAS

Costa, B. (2019). Introdução à Engenharia de Dados na AWS. Medium.
Link: https://medium.com/@bernardo.costa/introdu%C3%A7%C3%A3o-%C3%A0-engenharia-de-dados-na-aws-e3fd122ebb78

Costa, B. (2023). Como a AWS pode ajudar a coletar e processar grandes volumes de dados. Medium.
Link: https://medium.com/@bernardo.costa/como-a-aws-pode-ajudar-a-coletar-e-processar-grandes-volumes-de-dados-5e25ca3f858

Leafio.ai. (2020). Enfrentando os 5 principais desafios da previsão de demanda no varejo. Leafio AI.
Link: https://www.leafio.ai/pt/blog/enfrentando-os-5-principais-desafios-da-previsao-de-demanda-no-varejo/

Optidata. (2021). Como a tecnologia Cloud pode ajudar o varejo a crescer. Optidata.
Link: https://www.optidata.cloud/como-a-tecnologia-cloud-pode-ajudar-o-varejo-a-crescer/

Brascloud. (2020). A importância do Cloud Computing para o varejo. Brascloud.
Link: https://www.brascloud.com.br/pt-br/blog/a-importancia-do-cloud-computing-para-o-varejo

Tecnicon. (2020). Machine Learning no controle de estoque: você sabe como aplicar? Tecnicon.
Link: https://www.tecnicon.com.br/blog/479-Machine_Learning_no_controle_de_estoque_voce_sabe_como_aplicar_

Amazon Web Services. (n.d.). Arquitetura AWS para empresas: Soluções e metodologias. Amazon Web Services.
Link: https://aws.amazon.com/pt/architecture/?cards-all.sort-by=item.additionalFields.sortDate&cards-all.sort-order=desc&awsf.content-type=*all&awsf.methodology=*all&awsf.tech-category=*all&awsf.industries=*all&awsf.business-category=*all