Projeto Integrador IV - Translog PB

César Santos Nuevo Galan - 2041382311003
Felipe da Silva Santos Moreira - 2041382311007
William Ferreira dos Santos - 2041382311033

Relatório Técnico-Científico apresentado na disciplina de Projeto Integrador para o curso de Big Data para Negócios da Fatec Ipiranga “Pastor Enéas Tognini”. Desenvolvido para Translog Pombinha Branca para comprovação das atividades extensionistas.

São Paulo - SP
2024

RESUMO

Este projeto de conclusão semestral foca no desenvolvimento e implementação de uma solução de análise de dados para a área de transporte de produtos químicos, com ênfase na criação de um Data Lake e Data Lakehouse, além da aplicação de ferramentas de Business Intelligence e Inteligência Artificial para otimizar rotas, prever custos e monitorar a performance das operações. O estudo incluiu um levantamento detalhado sobre a empresa Translog Pombinha Branca, seu mercado de atuação, e as tecnologias utilizadas, complementado por uma análise bibliográfica.

O objetivo principal do estudo é demonstrar como as tecnologias de análise de dados podem melhorar a eficiência logística e a tomada de decisões estratégicas, aplicando modelos preditivos e soluções de BI para prever custos futuros de transporte e otimizar a gestão de rotas. O trabalho também aborda a criação de um pipeline de dados automatizado de ponta a ponta, que permite a ingestão, processamento e análise de dados históricos e em batch de forma contínua. Este pipeline visa integrar dados de diferentes fontes e transformar essas informações em insights acionáveis por meio de dashboards operacionais e financeiros. Além disso, a integração de modelos de Inteligência Artificial para previsão de custos de combustíveis é um componente central da solução. O objetivo geral é aprimorar a gestão das operações de transporte através da análise de dados, promovendo a otimização e geração de insights que ajudem na melhoria contínua dos serviços da empresa.

ABSTRACT

This semester-end project focuses on the development and implementation of a data analysis solution for the chemical products transportation industry, with an emphasis on creating a Data Lake and Data Lakehouse, as well as applying Business Intelligence and Artificial Intelligence tools to optimize routes, predict costs, and monitor operational performance. The study included a detailed survey of information about Translog Pombinha Branca, its market, and the technologies used, complemented by a bibliographic analysis.

The main objective of the study is to demonstrate how data analysis technologies can improve logistical efficiency and strategic decision-making by applying predictive models and BI solutions to forecast transportation costs and optimize route management. The project also involves the creation of an automated end-to-end data pipeline, which facilitates the continuous ingestion, processing, and analysis of historical and batch data. This pipeline integrates data from various sources and transforms it into actionable insights through operational and financial dashboards. Additionally, the integration of Artificial Intelligence models for fuel cost forecasting is a central part of the solution. The overall goal is to enhance transportation management through data analysis, optimizing operations, and generating insights that support the continuous improvement of the company's services.

SUMÁRIO

1 Introdução

O setor de transporte de produtos químicos exige soluções logísticas especializadas, dadas as particularidades relacionadas à segurança, conformidade regulatória e rastreabilidade das cargas. Neste contexto, a Translog Pombinha Branca destaca-se como uma microempresa localizada em São Paulo, dedicada ao transporte urbano de cargas industrializadas, com ênfase no atendimento a clientes corporativos.

Fundada em 2020, a empresa atua exclusivamente no estado de São Paulo, oferecendo serviços que priorizam segurança, pontualidade e excelência no atendimento. Sua especialização no transporte de produtos químicos reflete o compromisso em garantir operações logísticas eficientes e conformes com as normativas vigentes, atendendo às exigências de um mercado dinâmico e competitivo.

Este trabalho explora como a aplicação de tecnologias avançadas, como Business Intelligence, Inteligência Artificial e Engenharia de Dados, pode contribuir para a otimização das operações da empresa. O foco está na integração de dados, automação de processos e criação de insights estratégicos que possibilitem maior eficiência, redução de custos e melhoria contínua nos serviços oferecidos pela empresa.

2 Desenvolvimento

2.1 Objetivos

Este projeto foca no desenvolvimento e implementação de uma solução de análise de dados para a área de transporte de produtos químicos, com ênfase na criação de um Data Lake e Lakehouse, além da aplicação de ferramentas de Business Intelligence e Inteligência Artificial para otimizar rotas, prever custos e monitorar a performance das operações. O estudo incluiu um levantamento detalhado sobre a empresa Translog Pombinha Branca, seu mercado de atuação, e as tecnologias utilizadas, complementado por uma análise bibliográfica.

O objetivo principal do estudo é demonstrar como as tecnologias de análise de dados podem melhorar a eficiência logística e a tomada de decisões estratégicas, aplicando modelos preditivos e soluções de BI para prever custos futuros de transporte e otimizar a gestão de rotas. O trabalho também aborda a criação de um pipeline de dados automatizado de ponta a ponta, que permite a ingestão, processamento e análise de dados históricos e em batch de forma contínua. Este pipeline visa integrar dados de diferentes fontes e transformar essas informações em insights acionáveis por meio de dashboards operacionais e financeiros. Além disso, a integração de modelos de Inteligência Artificial para previsão de custos de combustíveis é um componente central da solução. O objetivo geral é aprimorar a gestão das operações de transporte através da análise de dados, promovendo a otimização e geração de insights que ajudem na melhoria contínua dos serviços da empresa.

2.1.1 Definição do Negócio

A Translog Pombinha Branca é uma microempresa de transporte localizada em São Paulo, especializada em soluções logísticas e transporte de cargas urbanas. A empresa atua exclusivamente no estado de São Paulo, oferecendo serviços de entrega de produtos industrializados para clientes corporativos. Com foco em segurança, pontualidade e excelência no atendimento, a Translog Pombinha Branca é especializada no transporte de produtos químicos, garantindo soluções logísticas seguras e eficientes para esse segmento específico.

2.1.2 A Empresa

Comprovante Situação Cadastral

Razão Social: LUCIVAL ALVES DOS SANTOS 58387862568
Nome Fantasia: Translog Pombinha Branca
Porte: Microempresa
CNPJ: 09.446.165/0001-66
Data de Abertura: 06/05/2020
Site (em construção): https://www.translogpombinhabranca.com/
Responsável: Lúcio Santos
Email: lucio.santos@translogpombinhabranca.com
Atividade: Transporte Rodoviário de Carga

Anexo 1 - Logotipo da empresa. Fonte: Translog Pombinha Branca

2.1.3 Pesquisa de Mercado

O mercado de transporte de produtos químicos no estado de São Paulo apresenta particularidades importantes em 2024. Este segmento enfrenta desafios como a crescente demanda por conformidade regulatória, segurança operacional e otimização logística. Ao mesmo tempo, oferece oportunidades significativas devido à alta demanda por transporte especializado em setores industriais e manufatureiros.

2.1.4 Tendências e Mudanças no Setor de Transporte de Produtos Químicos

- **Regulamentações Rigorosas:** A legislação ambiental e de transporte seguro exige que empresas invistam em processos e equipamentos que garantam o manuseio adequado de produtos perigosos.
- **Tecnologia e Rastreabilidade:** O uso de sistemas de rastreamento em tempo real está se tornando essencial para monitorar cargas, garantir segurança e otimizar rotas.
- **Sustentabilidade:** Empresas que adotam práticas sustentáveis, como otimização de rotas e veículos menos poluentes, conquistam vantagens competitivas.
- **Parcerias Estratégicas**: As indústrias químicas buscam transportadoras confiáveis para minimizar riscos e maximizar a eficiência..
- **Concorrência e Mercado Local:** Embora existam grandes operadores logísticos no estado, empresas especializadas como a Translog Pombinha Branca têm vantagem em oferecer um serviço altamente personalizado e focado nas necessidades específicas de transporte de produtos químicos.

2.1.3 Missão, Visão e Valores

Missão: Garantir o transporte seguro, eficiente, com foco em atendimento personalizado e excelência operacional.

Visão: Ser reconhecida como líder em transporte no estado de São Paulo, destacando-se pela segurança, pontualidade e inovação.

Valores:

- **Segurança:** Prioridade máxima na manipulação e transporte de cargas.
- **Eficiência:** Compromisso em entregar soluções ágeis e confiáveis.
- **Responsabilidade Ambiental:** Minimizar impactos no meio ambiente.

2.1.4 Requisitos

- **Modelos de Previsão de Custos e Receitas:** Simulações baseadas em dados históricos para estimar lucratividade por rota e cliente.
- **Dashboards Operacionais e Financeiros:** Visualizações que facilitem o monitoramento de KPIs críticos, como custo por km e eficiência dos motoristas, com dados atualizados em batch.
- **Segmentação de Clientes:** Análise para identificar os principais clientes e criar estratégias para atender às suas demandas.
- **Planejamento de Rotas com Algoritmos de Otimização:** Redução de custos e tempo de entrega por meio de análise preditiva e algoritmos como Dijkstra.

2.2 Justificativa e delimitação do problema

O problema que norteia este estudo pode ser formulado na seguinte pergunta: *"Como a análise de dados pode contribuir para a eficiência logística, a redução de custos e a otimização de rotas no transporte de produtos químicos?"*

A proposta se justifica pela relevância do tema em um cenário onde empresas de transporte enfrentam desafios significativos, como regulamentações rigorosas, necessidade de rastreabilidade e exigências de sustentabilidade.

Este projeto oferece benefícios ao propor soluções que aumentam a eficiência logística, otimizam rotas e permitem a tomada de decisões baseadas em dados. Além disso, a aplicação de tecnologias modernas, como Inteligência Artificial e pipelines de dados automatizados, possibilita atender à crescente demanda por inovação no setor de transporte de produtos químicos, alinhando-se às melhores práticas de mercado e contribuindo para a competitividade da empresa.

2.3 Fundamentação teórica

2.3.1 Transporte de Produtos Químicos

A fundamentação teórica deste trabalho é baseada em fontes atuais, destacando-se os seguintes tópicos:

- **Transporte:** A segurança e eficiência no transporte de produtos químicos são reguladas por normas rigorosas, conforme descrito pela ABTLP (2024) e SETCESP (2024). Essas regulamentações incluem práticas obrigatórias para o manuseio seguro e rastreamento de cargas perigosas, um fator crucial para a competitividade no mercado.
- **Data Lake e Data Lakehouse:** Segundo estudos recentes, a adoção de Data Lakes e Data Lakehouses permite a ingestão e o armazenamento eficiente de dados estruturados e não estruturados, viabilizando análises em tempo real e em batch (KARTHIKEYAN, 2021).
- **Business Intelligence e Dashboards:** Ferramentas de BI como dashboards interativos têm se mostrado eficazes para monitorar KPIs críticos, como custo por quilômetro e eficiência operacional (POWER et al., 2018).
- **Inteligência Artificial e Previsões de Custos:** A aplicação de modelos preditivos para a estimativa de custos de combustíveis e planejamento de rotas está alinhada com tendências do setor, promovendo eficiência e redução de desperdícios (REVISTA EMPREENDE, 2024).

2.3.2 Data Lake, Data Warehousing e CSP

Uma **CSP (Cloud Service Provider)** é uma empresa terceirizada que oferece serviços de computação em nuvem sob demanda e escalonáveis, como armazenamento, processamento e redes, através da internet. Esses serviços permitem que as empresas e usuários acessem recursos de TI sem a necessidade de gerenciar a infraestrutura física, pagando apenas pelo que utilizam. Os CSPs como Google Cloud, AWS e Microsoft Azure oferecem uma variedade de modelos de serviço, incluindo IaaS, PaaS e SaaS, cada um atendendo a diferentes necessidades de negócios e desenvolvimento.

A computação em nuvem está se tornando o quadro preferido para impulsionar a transformação digital e a inovação, pois oferece flexibilidade, escalabilidade, resiliência e segurança necessárias para atender às demandas de negócios, sem as limitações dos servidores locais ou o alto custo de criar e manter um data center interno. Em vez de construir sua própria infraestrutura, o cliente pode alugar serviços de um CSP e compartilhá-los com outras empresas ou indivíduos. Com a escalabilidade, flexibilidade, economia de custos e segurança oferecidas pelos CSPs, as empresas conseguem aumentar sua eficiência operacional e inovar de forma mais rápida e eficaz.

Se tratando de soluções de armazenamento de dados, um Data Lake permite armazenar dados em sua forma bruta, sejam eles estruturados, semiestruturados ou não estruturados. Diferente dos bancos de dados tradicionais, um Data Lake mantém os dados no formato original até que sejam necessários para análise. Isso oferece flexibilidade para empresas analisarem e processarem esses dados posteriormente usando ferramentas avançadas de big data, machine learning e inteligência artificial. Data Lakes são fundamentais para organizações que precisam gerenciar e explorar grandes quantidades de dados de diversas fontes.

2.3.2.1 Montagem de um Data Lake

A criação de um data lake envolve uma série de etapas essenciais para garantir que ele seja eficiente, seguro e atenda às necessidades da organização. O primeiro passo na construção de um data lake é definir requisitos. Isso significa compreender claramente quais são os objetivos principais da organização ao criar o data lake. Para isso, é necessário identificar tanto as necessidades de negócios quanto as necessidades técnicas. A organização deve especificar os tipos de dados que serão armazenados

(estruturados, semi-estruturados e não estruturados), as fontes desses dados (como sistemas internos, APIs externas, dispositivos IoT, etc.), os requisitos de desempenho (como velocidade de ingestão e processamento de dados) e as exigências de segurança e conformidade (como GDPR, LGPD ou outras regulamentações regionais ou setoriais).

Uma vez que os requisitos estejam bem definidos, o próximo passo é projetar a arquitetura do data lake. Esta arquitetura deve ser robusta e escalável para acomodar o volume crescente de dados. Ao projetar a arquitetura, também é fundamental estabelecer diretrizes de segurança para proteger os dados armazenados no data lake. Isso envolve definir mecanismos de autenticação e autorização para garantir que apenas usuários e sistemas autorizados tenham acesso aos dados sensíveis. Além disso, a criptografia dos dados em trânsito e em repouso é essencial para proteger a integridade e a confidencialidade dos dados.

Outro aspecto crucial do projeto da arquitetura é a documentação e catalogação dos metadados. Os metadados fornecem contexto e informações detalhadas sobre os dados, facilitando sua descoberta e uso. Uma boa prática é usar ferramentas de catalogação de dados, que ajudam a manter o controle e a organização dos dados no data lake. Além disso, é essencial garantir que o data lake esteja em conformidade com as regulamentações aplicáveis, assegurando que as políticas de acesso e segurança atendam às exigências legais e éticas.

2.3.2.2 Escolha Para o Projeto

No projeto Translog Pombinha Branca, a escolha das ferramentas de armazenamento e processamento de dados foi baseada em critérios como escalabilidade, flexibilidade, custo-benefício e integração com tecnologias analíticas modernas. A combinação de um Data Lake na AWS S3, um Data Lakehouse e a plataforma Databricks oferece uma infraestrutura robusta para atender às necessidades de coleta, organização, transformação e análise dos dados.

O Amazon S3 foi escolhido como repositório base por sua capacidade de armazenar uma quantidade ilimitada de dados, acompanhando o crescimento contínuo do volume de informações. Sua estrutura é altamente custo-eficiente, uma vez que utiliza o modelo de pagamento por uso, tornando o armazenamento acessível mesmo para grandes volumes de dados. O suporte a formatos como Parquet é uma característica essencial, garantindo compressão eficiente e leitura rápida, aspectos fundamentais para o desempenho do pipeline. Além disso, a infraestrutura da AWS oferece alta disponibilidade e confiabilidade, assegurando que os dados estejam sempre acessíveis e protegidos.

No contexto do pipeline de dados da Translog Pombinha Branca, o bucket Raw - "*pb-translog-raw*" - da AWS S3 desempenha um papel crucial como repositório inicial de armazenamento, consolidando dados coletados de diversas fontes antes de serem

processados nas etapas subsequentes do pipeline. Ele é utilizado como a base do Data Lake, armazenando os dados em seu estado bruto e garantindo sua preservação para análises futuras.

Todos os dados extraídos no pipeline no bucket Raw que é configurado para utilizar o formato Parquet, conhecido por sua eficiência na compressão e leitura, o que otimiza tanto o uso do espaço de armazenamento quanto o tempo necessário para processamento dos dados.

No contexto do pipeline, o bucket RAW serve como ponto inicial de ingestão e é essencialmente o Data Lake do projeto. Ele atua como a principal fonte de dados para as camadas subsequentes Bronze, Silver e Gold, que são gerenciadas e transformadas no Databricks. Os dados coletados automaticamente pelo Zapier, especialmente aqueles extraídos de e-mails, também são armazenados diretamente nesse bucket, consolidando-o como a base de todas as operações de processamento e análise de dados no sistema.

2.3.3 Construção de Dashboards utilizando Análise Estatística e Machine Learning

Os dashboards desenvolvidos para o Projeto Integrador desempenham um papel essencial na transformação de dados em informações valiosas e visualmente acessíveis. Esses painéis foram criados com o objetivo de fornecer insights estratégicos que apoiam a tomada de decisão e otimizam os processos operacionais da empresa.

Cada dashboard foi planejado para atender a áreas específicas da organização, possibilitando a análise detalhada e personalizada de métricas críticas para o negócio. As visualizações foram elaboradas com foco na clareza e na eficiência, garantindo que os stakeholders possam identificar padrões, acompanhar KPIs e identificar oportunidades de melhoria com facilidade.

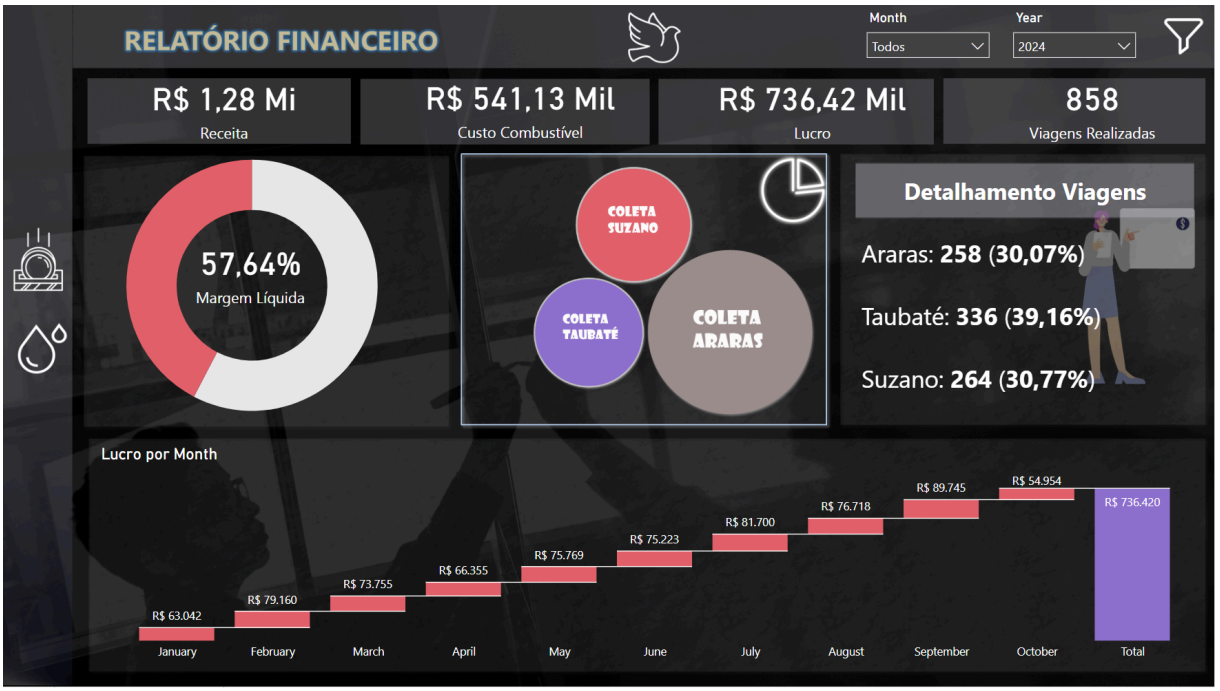
Por meio do uso da ferramenta avançada de visualização Power BI, os dashboards consolidam dados brutos, previamente processados em diferentes camadas do pipeline de dados, em representações claras e impactantes. Essa abordagem não apenas aumenta a transparência, mas também promove maior confiança na análise, dado o rigor com que os dados foram tratados ao longo do processo.

No âmbito do Machine Learning do projeto, foi desenvolvido um algoritmo de previsão ARIMA (AutoRegressive Integrated Moving Average), que é um modelo estatístico amplamente utilizado para análise e previsão de séries temporais. Ele combina componentes autorregressivos (AR), de média móvel (MA) e de diferenciação

para lidar com séries que não são estacionárias. O modelo é especialmente útil para detectar padrões e realizar previsões com base em dados históricos.

Para exemplificar o uso do ARIMA, foi utilizado o dataset disponibilizado pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), que contém a série histórica de preços de combustíveis no Brasil. Esse conjunto de dados inclui informações sobre preços de combustíveis automotivos, GLP, óleo diesel (S-500 e S-10), etanol e gasolina comum. Ele pode ser acessado na página de Dados Abertos do governo brasileiro, no link fornecido, e é uma excelente fonte para estudos relacionados ao mercado de combustíveis.

Uma aplicação prática seria carregar o dataset, realizar a análise exploratória dos dados para compreender padrões de variação nos preços ao longo do tempo, e então aplicar o ARIMA para prever preços futuros. Isso pode ser útil para empresas de transporte, políticas públicas e consumidores em geral, ajudando a antecipar flutuações de preços e a planejar orçamentos ou políticas de intervenção.



Anexo 2 - Dashboard de Relatório Financeiro. Fonte: Autores

Analisando o dashboard apresentado, é possível identificar as seguintes métricas:

1. **Receita Total**
2. **Custo com Combustível**

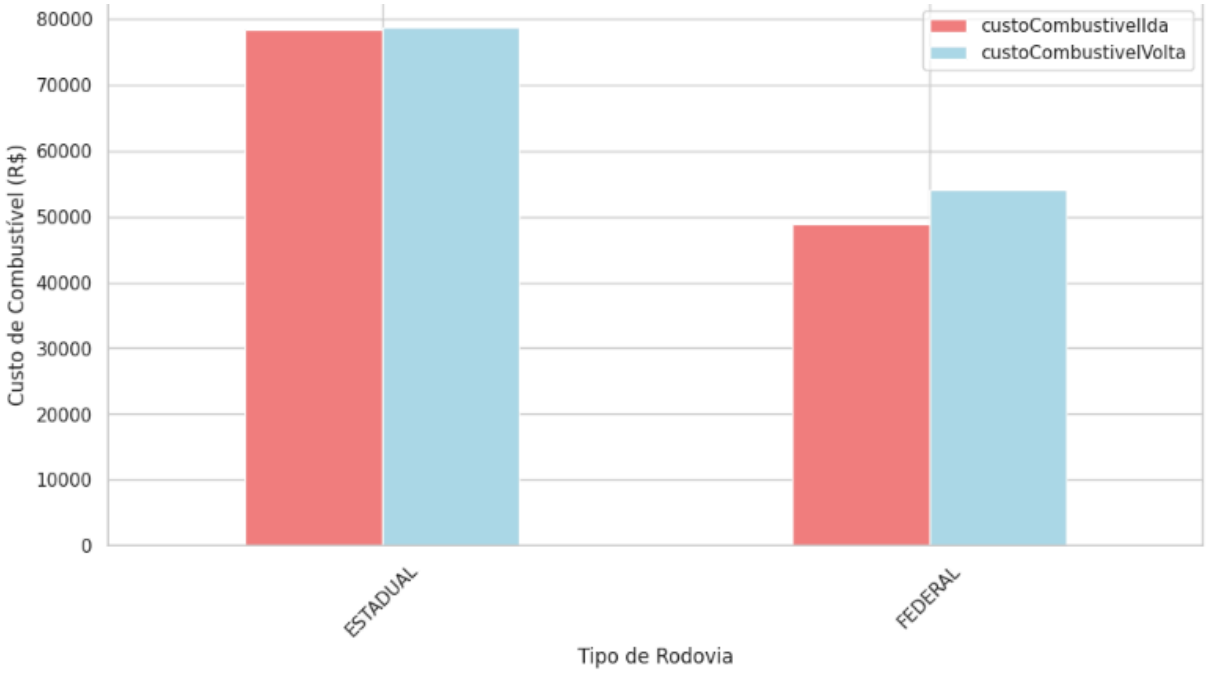
3. **Lucro Total**
4. **Viagens Realizadas**
5. **Margem Líquida**
6. **Distribuição de Coletas por Cidade**
7. **Lucro por Mês**

Anexo 3 - Dashboard de Relatório de Consumo. Fonte: Autores

Segue a lista de métricas do relatório de consumo exibido na imagem. A "Previsão Diesel" foi gerada através de uma previsão realizada com o algoritmo de Machine Learning ARIMA, que estima os custos do combustível diesel utilizado pelos veículos de carga da Translog Pombinha Branca.

1. **Custo com Combustível**
2. **Média de Combustível por Viagem**
3. **Litros Consumidos**
4. **Custo de Combustível por Mês**
5. **Distribuição de Coletas por Cidade**
6. **Previsão Diesel**

É possível conceber insights através de métricas simples de estatística através do uso de ferramentas como Pandas e Numpy, além de bibliotecas gráficas como Pyplot e Seaborn. Vejamos abaixo um exemplo:

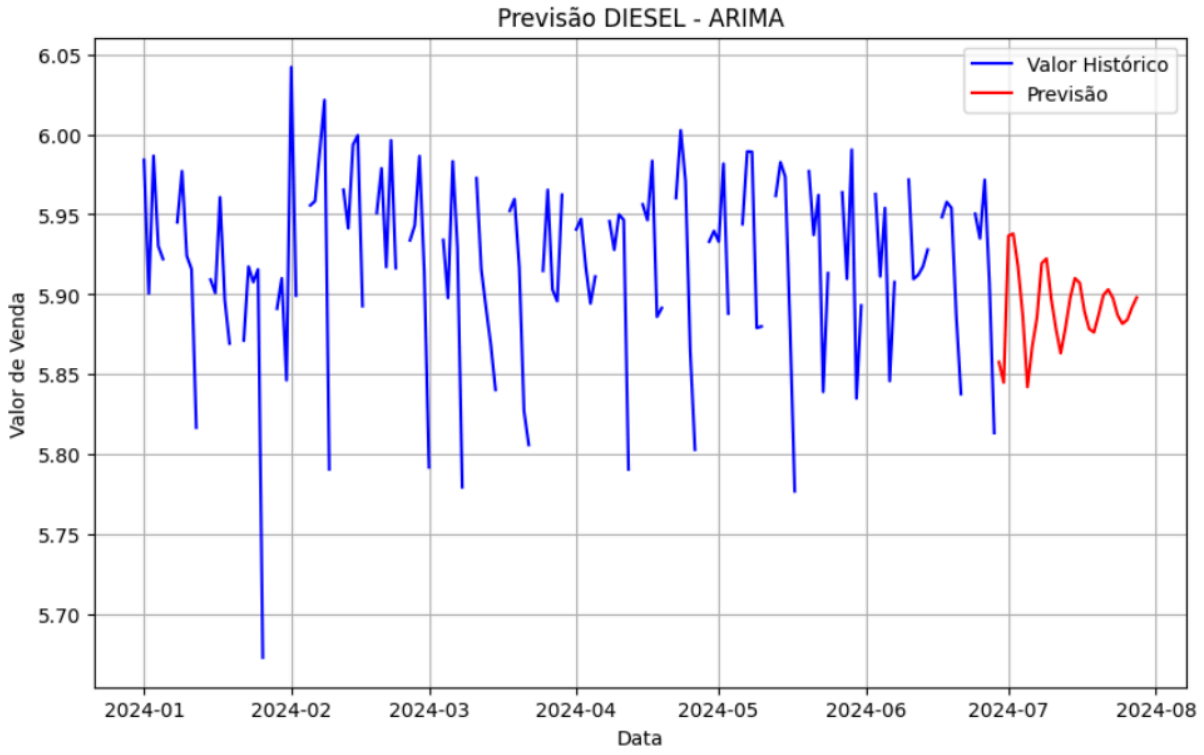


Anexo 4 - Gráfico de Quantidade de Cidades Visitadas. Fonte: Autores

Um dos objetivos macros do trabalho é otimizar a escolha de rotas para gerar economias e, consequentemente, lucros. No gráfico acima, é possível observar a diferença nos preços praticados por tipos de rodovia. A análise foi realizada considerando o peso transportado pelo caminhão e o preço dos combustíveis nos respectivos dias.

2.3.3.1 Previsões com Arima

O modelo testado no Projeto Integrador conseguiu gerar o seguinte resultado abaixo para o combustível do tipo Diesel:



Anexo 6 - Gráfico de Previsões do modelo ARIMA. Fonte: Autores

Baseando-se no histórico de preços reportados, o modelo conseguiu ser satisfatório ao gerar padrões que sinalizam tendências de aumento de preço em dias determinados.

2.3.4 ETL E ELT

O ETL/ELT faz parte de um processo crucial em todo projeto de dados - os dados precisam passar por processos de transformação, filtragem e mecanismos para que as informações sensíveis sejam preservadas. Afinal, só é possível gerar valor quando os dados passam por seus devidos processos, integrações e caminhos até o usuário final.

Tradicionalmente, as empresas criam pipelines ETL/ELT através de programação manual, o que é ineficaz e demorado. O problema principal com a programação é que os dados são manipulados por elementos em vez de colunas e linhas, tornando a integração de fontes heterogêneas mais difícil. Além disso, modificar ou adicionar pipelines de dados requer a criação de bibliotecas anteriores e a integração de código complicado.

Mediante a tais fatos, as empresas modernas estão adotando ferramentas automatizadas que podem lidar com processos complexos de forma mais eficiente. Essas ferramentas simplificam, abstraindo complexidades e permitindo que usuários não-técnicos lidem com transformações de dados e fluxos de trabalho. Essa abstração acelera o desenvolvimento, a manutenção e a escalabilidade, tornando os processos ETL/ELT mais acessíveis e eficientes para uma variedade maior de usuários dentro de uma organização.

2.3.4.1 Comparação ETL x ELT

Grandes empresas possuem diversas fontes de dados em suas operações, como aplicações, sensores, infraestrutura de TI e parceiros terceirizados. É necessário filtrar, classificar e limpar esses dados para análise e inteligência de negócios. A abordagem ETL aplica regras de negócios para processar dados antes da integração centralizada, enquanto a abordagem ELT carrega os dados como estão e os transforma posteriormente. O processo de ETL exige definição inicial e envolvimento de analistas para determinar tipos de dados, estruturas e relações. Cientistas de dados utilizam o ETL para carregar dados em data warehouses, sendo a ELT a prática mais comum atualmente.

O processo de ETL surgiu nos anos 70 e ganhou destaque com o advento dos data warehouses. Antigamente, os data warehouses demandavam procedimentos personalizados de ETL para cada fonte de informação. Com o avanço das tecnologias em nuvem, houve uma mudança significativa. Agora, as empresas podem armazenar uma quantidade ilimitada de dados brutos em larga escala e analisá-los posteriormente, conforme necessário. O ELT se tornou o método contemporâneo de integração de dados para análises eficazes.

2.3.4.2 Python e Spark

No contexto de pipelines de dados, os processos ETL (Extract, Transform, Load) e ELT (Extract, Load, Transform) são fundamentais para mover, processar e transformar

grandes volumes de informações. Eles têm como objetivo transformar dados brutos de diversas fontes em dados organizados e prontos para análises e consumo por ferramentas de BI.

Com ferramentas como Apache Spark e linguagens como Python, esses processos podem ser implementados de forma eficiente, especialmente em ambientes que requerem escalabilidade e processamento distribuído. O ETL realiza o fluxo de dados em três etapas sequenciais:

- **Extract (Extração):** Dados são extraídos de sistemas de origem como bancos de dados relacionais, APIs, arquivos CSV ou sites. Em Python, bibliotecas como *pandas*, *pyodbc*, ou conectores específicos são utilizadas para acessar e extrair os dados. No Spark, a extração é feita por meio do módulo *Spark SQL*, permitindo conexões a diversas fontes, como JDBC e S3.
- **Transform (Transformação):** Dados extraídos são processados para normalização, agregação, enriquecimento ou limpeza. O Python, usa ferramentas como *pandas*, enquanto no Spark, por sua vez, utiliza o *DataFrame API* e o *Spark SQL* para operações em larga escala.
- **Load (Carregamento):** Os dados transformados são carregados em um destino como data warehouses, data lakes ou bancos de dados analíticos. O Spark utiliza conectores como o *JDBC* para carregar dados diretamente nos destinos, enquanto Python pode usar bibliotecas como *SQLAlchemy* para a mesma finalidade.

No modelo ELT, a sequência é modificada:

- **Extract (Extração):** Igual ao ETL, os dados são retirados das fontes originais.
- **Load (Carregamento):** Os dados brutos são carregados diretamente na camada de armazenamento do data lake ou warehouse, como um bucket S3 ou GCS. O Spark facilita isso com suporte nativo a armazenamentos distribuídos.
- **Transform (Transformação):** As transformações são realizadas diretamente no destino (como um Delta Lake ou Snowflake), aproveitando o poder de processamento dessas plataformas.O Spark é amplamente utilizado para transformação por meio de scripts PySpark, especialmente em fluxos baseados no formato Delta Lake, permitindo versionamento e transações ACID.

2.3.4.3 Base de Dados

Para análise inicial, o time de desenvolvedores optou por utilizar o dataset "Delivery truck trips data" obtido no Kaggle. Com base nessa base de dados, as Regras de Negócio são as diretrizes que definirão como a Translog Pombinha Branca gerencia seus processos de transporte e logística, otimizando operações e maximizando a eficiência com base em análises de dados. Essas regras estão alinhadas com as necessidades de entregas pontuais, utilização eficiente dos veículos e satisfação dos clientes.

A partir do dataset, as seguintes perguntas podem ser propostas para o cliente como base para análises e insights:

1. Quais rotas apresentam maior atraso e como otimizá-las?
2. Qual o impacto das condições geográficas no tempo de viagem?
3. Consumo de combustível e seu impacto na receita final.
4. Como os atrasos estão distribuídos em relação aos diferentes fornecedores/ clientes.
5. Qual a distância média percorrida por dia comparada ao mínimo esperado ?

2.4 Estratégias para o Pipeline de Dados

No desenvolvimento do pipeline de dados da *Translog Pombinha Branca*, utilizamos uma combinação estratégica de ferramentas e tecnologias para garantir eficiência, escalabilidade e integração entre as camadas do Data Lakehouse. A seguir, apresentamos as principais abordagens adotadas:

2.4.1 Automação de Extração com Zapier

O Zapier foi configurado para monitorar caixas de e-mail da empresa e identificar datasets enviados como anexos em planilhas. Quando detectados, os arquivos são automaticamente transferidos para a camada raw no S3 - bucket "*pb-translog-raw*" - garantindo que nenhuma informação seja perdida e possibilitando a automação no envio dos dados para o pipeline, reduzindo intervenções manuais.

2.4.2 Orquestração dos Dados entre Camadas do Data Lakehouse

A movimentação e transformação de dados entre as camadas Bronze, Silver e Gold são realizadas por meio de scripts Spark, executados no Databricks. Esse processo inclui:

- **Bronze**: Dados brutos carregados diretamente do S3.
- **Silver**: Limpeza, deduplicação e enriquecimento dos dados.
- **Gold**: Dados transformados e modelados, prontos para análises e visualizações.

Utilizamos o Delta Lake no Databricks para gerenciar versões de dados e garantir a consistência e rastreabilidade das transformações.

2.4.3 Conexão do DataBricks com Power BI

Os dashboards de análise foram desenvolvidos no Power BI, conectando-se diretamente às tabelas da camada Gold do Databricks. Essa integração permite atualizações de métricas e insights estratégicos para a empresa.

A conexão foi realizada utilizando a ferramenta Partner Connect do Databricks, garantindo uma integração fluida entre o Data Lakehouse e as ferramentas de visualização.

A combinação dessas tecnologias permite à **Translog Pombinha Branca** coletar, processar e visualizar dados de forma automatizada e escalável, alinhando-se às melhores práticas de gestão de dados para tomada de decisões baseada em insights. A abordagem adotada assegura flexibilidade para incorporar novas fontes de dados e escalar o pipeline conforme as necessidades do negócio evoluem.

2.5 Conclusão de Aplicação das Disciplinas Estudadas no Projeto Integrador

2.5.1 Gestão Econômica e Financeira

Os conceitos de análise financeira aprendidos nesta disciplina foram aplicados para calcular e projetar custos operacionais relacionados às rotas de transporte, além de avaliar a viabilidade econômica das estratégias propostas. Também foram utilizadas métricas como custo por quilômetro e lucro por cliente, que foram incorporadas nos

dashboards desenvolvidos, permitindo uma visualização clara e objetiva do desempenho financeiro da empresa.

2.5.2 Arquitetura de Big Data e DW/BI

Os conhecimentos adquiridos sobre arquiteturas de dados foram fundamentais para o projeto. Utilizando Databricks, foi possível criar um pipeline automatizado que integra dados de múltiplas fontes no Data Lake e transforma esses dados em insights no Data Lakehouse. As práticas de ETL e o design do ambiente para suportar análise em batch garantiram o fluxo contínuo e eficiente de informações para o sistema de BI.

2.5.3 Big Data Analytics I - Modelagem e Análise Estatística

A análise de dados foi realizada utilizando ferramentas do Databricks e bibliotecas Python. Os dados históricos foram modelados e analisados para identificar tendências e padrões, com os resultados visualizados em gráficos que facilitaram a tomada de decisões estratégicas. Essa abordagem foi crucial para compreender os fatores que impactam os custos de transporte.

2.5.4 Programação em Banco de Dados II

O ecossistema Hadoop foi empregado para lidar com grandes volumes de dados, permitindo a criação de um ambiente robusto para armazenamento e processamento. O conhecimento adquirido nesta disciplina foi essencial para integrar o Hadoop ao pipeline de dados, garantindo escalabilidade e eficiência no gerenciamento das informações.

2.5.5 Laboratório de Programação II

As bibliotecas Python, foram amplamente utilizadas para manipulação, análise e visualização dos dados. Essas ferramentas facilitaram a criação de relatórios e gráficos detalhados, além de apoiar a validação e o refinamento dos modelos preditivos desenvolvidos no projeto.

2.5.6 Aprendizagem de Máquina

O modelo ARIMA foi aplicado para realizar previsões precisas de custos

operacionais, como os relacionados ao consumo de combustível. O aprendizado dessa disciplina permitiu a escolha do modelo mais adequado e sua configuração para atender às necessidades do projeto, garantindo previsões confiáveis para a tomada de decisões estratégicas.

2.6 Metodologia

A metodologia utilizada no projeto seguiu as etapas "Ouvir, Criar e Implementar" recomendadas pela UNIVESP, empregando métodos específicos em cada fase para garantir a execução bem-sucedida do projeto.

O contexto foi analisado por meio de entrevistas com o responsável pela Translog Pombinha Branca e coleta de informações operacionais da empresa. Dados foram levantados utilizando observação direta e análises qualitativas sobre os desafios enfrentados, como ineficiências nas rotas e dificuldade na previsão de custos.

A prototipação foi aplicada através de análises quantitativas nos dados coletados, utilizando ferramentas do Databricks e bibliotecas Python como Pandas para identificar padrões e tendências. A solução desenvolvida incluiu a criação de um pipeline de dados automatizado baseado no ecossistema Hadoop, um dashboard interativo para monitoramento de KPIs e um modelo preditivo ARIMA, que gerou estimativas precisas de custos operacionais. Diagramas e modelos foram usados para validar a arquitetura do Data Lakehouse e detalhar o fluxo de ingestão e transformação dos dados.

Esse conjunto estruturado de estratégias e metodologias possibilitou alcançar os objetivos do projeto de forma eficaz, garantindo que as soluções desenvolvidas fossem não apenas aplicáveis à realidade da empresa, mas também capazes de gerar impacto significativo em suas futuras operações logísticas.

3 Resultados: solução final

Identificando os principais desafios do projeto, o grupo criou um sistema funcional composto por um pipeline de dados automatizado para ingestão e processamento de informações históricas e em tempo real, um dashboard interativo que permite monitorar KPIs operacionais e financeiros, e um modelo preditivo de Inteligência Artificial para estimar custos com combustíveis e sugerir rotas otimizadas.

O produto final integrou as soluções descritas, resultando em um sistema completo para gestão de dados e otimização de operações logísticas. A solução não apenas atendeu às necessidades identificadas, mas também apresentou potencial de escalabilidade para novas áreas de atuação da empresa. Como resultado, proporcionou insights estratégicos para a empresa e contribuiu diretamente para a melhoria contínua de suas operações logísticas.

4 Considerações finais

O trabalho apresentado teve como objetivo desenvolver uma solução de análise de dados para a empresa Translog Pombinha Branca, especializada no transporte de substâncias químicas no estado de São Paulo. A proposta respondeu aos desafios específicos do setor, como a necessidade de otimizar rotas, prever custos operacionais e monitorar o desempenho operacional, com base em tecnologias como Business Intelligence, Inteligência Artificial e arquitetura de dados Data Lake e Data Lakehouse.

As contribuições do trabalho incluem fornecer uma solução que reduza custos e aprimore o planejamento logístico da empresa. Os resultados previstos reforçam a importância da aplicação de tecnologias de rastreabilidade e otimização logística. Além disso, o uso de tecnologias avançadas mostrou como as pequenas empresas podem utilizar a análise de dados para ganhar concorrência no mercado. Porém, foram identificadas limitações, como a necessidade de maior maturidade tecnológica da organização para adoção plena da solução, bem como desafios relacionados à integração de dados em tempo real.

A adoção de soluções de análise de dados para promover a melhoria contínua dos serviços prestados é de suma importância para qualquer empresa. A redução de custos, o aumento da eficiência e a maior transparência das operações reforçam o valor da tecnologia como ferramenta de sustentabilidade e crescimento do setor logístico. Esses avanços ajudam a fortalecer o mercado local, promover práticas inovadoras e demonstrar a viabilidade de soluções tecnológicas para pequenos negócios.

Referências

ABNT – Associação Brasileira de Normas Técnicas. **NBR 14724**: Informação e documentação. Trabalhos Acadêmicos - Apresentação**.** Rio de Janeiro: ABNT, 2002.

ABTLP. Transporte Químico. Disponível em: https://www.abtlp.org.br/index.php/transporte-quimico/. Acesso em: 27 agosto de 2024. 07:50

SETCESP. ANTT atualiza regulamento para o transporte de produtos perigosos. Disponível em: https://setcesp.org.br/noticias/legislacao/antt-atualiza-regulamento-para-o-transporte-de-produtos-perigosos/. Acesso em: 28 agosto de 2024. 07:50

REVISTA EMPREENDE. Transporte rodoviário de produtos químicos vislumbra ano promissor em 2024. Disponível em: https://revistaempreende.com.br/transporte-rodoviario-de-produtos-quimicos-vislumbra-ano-promissor-em-2024/. Acesso em: 29 de novembro de 2024. 17:00.

BLOG DO CAMINHONEIRO. Papo de Papelaria. Disponível em: https://blogdocaminhoneiro.com/2024/01/transporte-rodoviario-de-produtos-quimicos-vislumbra-ano-promissor-em-2024/. Acesso em: 29 de novembro de 2024. 07:30.

AWS. Armazenamento de objetos na nuvem – Amazon S3. Disponível em: https://aws.amazon.com/pt/pm/serv-s3/?gclid=CjwKCAjwlbu2BhA3EiwA3yXyuz2lNP6fz2QNgGNa_xCAdH4jzRawIT6B_d2j4NvZxyy7c2aG3yQNoRoCa9UQAvD_BwE&trk=9c7f9c59-8d98-452d-8a14-441a9b6492f3&sc_channel=ps&ef_id=CjwKCAjwlbu2BhA3EiwA3yXyuz2lNP6fz2QNgGNa_xCAdH4jzRawIT6B_d2j4NvZxyy7c2aG3yQNoRoCa9UQAvD_BwE:G:s&s_kwcid=AL!4422!3!589951433465!e!!g!!amazon%20s3!16393976584!133547553013. Acesso em: 28 de agosto de 2024. 08:00.

GOOGLE. O que é Apache Hadoop. Disponível em: https://cloud.google.com/learn/what-is-hadoop?hl=pt-BR. Acesso em: 26 de agosto de 2024.19:00.

BITWISE. Traditional ETL vs ELT on Hadoop. Disponível em: https://www.bitwiseglobal.com/en-us/traditional-etl-vs-elt-on-hadoop/. Acesso em 11 de setembro de 2024. 18:10.

KAGGLE. Delivery truck trips. Disponível em: https://www.kaggle.com/datasets/ramakrishnanthiyagu/delivery-truck-trips-data. Acesso em 06 de setembro de 2024. 21:10.

APACHE SPARK. Apache Spark - A Unified engine for large-scale data analytics. Disponível em: https://www.bitwiseglobal.com/en-us/traditional-etl-vs-elt-on-hadoop/. Acesso em 06 de setembro de 2024. 22:00.

DATABRICKS. Databricks on AWS. Disponível em: https://docs.databricks.com/en/index.html. Acesso em 10 de setembro de 2024. 21:10.

APACHE SPARK. Apache Spark - A Unified engine for large-scale data analytics. Disponível em: https://www.bitwiseglobal.com/en-us/traditional-etl-vs-elt-on-hadoop/. Acesso em 10 de setembro de 2024. 22:30.

ZAPIER. Gmail Integrations. Disponível em: https://www.bitwiseglobal.com/en-us/traditional-etl-vs-elt-on-hadoop/. Acesso em 11 de setembro de 2024. 07:00.

QLIK. Data Lakes vs. Data Warehouses Comparison Guide. Disponível em: https://www.qlik.com/us/data-lake/data-lake-vs-data-warehouse. Acesso em 14 de setembro de 2024. 21:10.

DELTA LAKE. Welcome to the Delta Lake documentation. Disponível em: https://datasciencedojo.com/blog/data-lake-vs-data-warehouse/. Acesso em 15 de setembro de 2024. 20:00.

DATASCIENCEDOJO. Data Lakes vs. Data Warehouses: Decoding the Data Storage Debate. Disponível em: https://datasciencedojo.com/blog/data-lake-vs-data-warehouse/. Acesso em 17 de setembro de 2024. 20:30.

AWS. What is Amazon S3?. Disponível em: https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html. Acesso em 17 de setembro de 2024. 22:30.

GOV.BR. Série Histórica de Preços de Combustíveis e de GLP. Disponível em: https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/serie-historica-de-precos-de-combustiveis. Acesso em 15 de novembro de 2024. 21:10.

MICROSOFT. Documentação do Power BI. Disponível em: https://learn.microsoft.com/pt-br/power-bi/. Acesso em 28 de novembro de 2024. 22:00.