INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

CAMPUS MONTEIRO

TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

CARLOS EDUARDO ALVES DE MELO JÚNIOR

RELATÓRIO DE ESTÁGIO SUPERVISIONADO EM ENGENHARIA DE DADOS NA EMPRESA COMPASS UOL

MONTEIRO-PB

2024

CARLOS EDUARDO ALVES DE MELO JÚNIOR

RELATÓRIO DE ESTÁGIO SUPERVISIONADO EM ENGENHARIA DE DADOS NA EMPRESA COMPASS UOL

Relatório de Estágio apresentado à Coordenação de Estágio do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Campus Monteiro, como requisito parcial para conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Me. Wanderley Almeida de Melo Júnior.

MONTEIRO-PB

2024

Dados Internacionais de Catalogação na Publicação – CIP
Bibliotecária responsável Porcina Formiga dos Santos Salgado CRB15/204
IFPB Campus Monteiro.

M528r Melo Junior, Carlos Eduardo Alves de.
Relatório de estágio supervisionado em Engenharia de Dados na Empresa COMPASS-UOL / Carlos Eduardo Alves de Melo Junior – Monteiro-PB. 2024.
35fls. : il.

Relatório (Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - IFPB campus, Monteiro.

Orientador: Prof. Me. Wanderley Almeida de Melo Junior.

1. Dados - engenharia - ferramentas 2. AWS 3. Empresa COMPASS UOL – Tecnologia I. Titulo .

CDU 004.422.6

CARLOS EDUARDO ALVES DE MELO JÚNIOR

RELATÓRIO DE ESTÁGIO SUPERVISIONADO EM ENGENHARIA DE DADOS NA EMPRESA COMPASS UOL

Relatório de Estágio apresentado à Coordenação de Estágio do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Campus Monteiro, como requisito parcial para conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas.

Aprovado em 13 de agosto de 2024.

BANCA EXAMINADORA

gov.br
Documento assinado digitalmente
WANDERLEY ALMEIDA DE MELO JUNIOR
Data: 10/09/2024 14:40:39-0300
Verifique em https://validar.iti.gov.br

Prof. Me. Wanderley Almeida de Melo Júnior (Orientador - IFPB)

gov.br
Documento assinado digitalmente
GILVONALDO ALVES DA SILVA CAVALCANTI
Data: 22/08/2024 17:13:00-0300
Verifique em https://validar.iti.gov.br

Prof. Esp. Gilvonaldo Alves da Silva Cavalcanti (Examinador - IFPB)

Prof. Esp. Wagner de Oliveira Santos (Examinador - IFPB)

RESUMO

Este relatório descreve as principais atividades que realizei durante meu estágio em Engenharia de Dados na Compass UOL, uma empresa líder que fornece soluções de tecnologia em diversos setores, utilizando serviços da AWS (Amazon Web Services) em seu escopo. Durante o estágio, tive a oportunidade de mergulhar no ambiente corporativo, aplicando de forma prática alguns dos conhecimentos adquiridos ao longo do curso de Análise e Desenvolvimento de Sistemas. Ao longo desse período, enfrentei desafios simulados, representando cenários e problemas reais, nos quais desenvolvi e implementei soluções utilizando as ferramentas e serviços reais da AWS. Essa abordagem proporcionou uma experiência valiosa no uso eficaz das ferramentas de engenharia de dados em um contexto empresarial. Essa vivência foi fundamental para minha formação acadêmica e profissional, preparando-me de maneira sólida para uma atuação eficaz no mercado de trabalho.

Palavras-chave: engenharia de dados; estágio supervisionado; AWS; processamento de dados.

ABSTRACT

This report outlines the main activities I conducted during my Data Engineering internship at Compass UOL, a leading company providing technology solutions across various sectors, utilizing AWS (Amazon Web Services) services within its scope. Throughout the internship, I immersed myself in the corporate environment, applying practical aspects of the knowledge gained during my Analysis and Systems Development course. During this period, I engaged in simulated challenges, representing real-world scenarios and problems, for which I developed and implemented solutions using actual AWS tools and services. This approach provided invaluable practical experience in effectively utilizing data engineering tools within a corporate setting. This experience significantly enriched my academic and professional development, solidly preparing me for effective performance in the job market.

Keywords: data engineering; supervised internship; AWS; data processing.

IDENTIFICAÇÃO DO CAMPO DE ESTÁGIO

Identificação da Empresa:

Nome: Compass.Uol Tecnologia Ltda

Bairro: CENTRO

Endereço: R CORONEL CHICUTA, 575, CONJ 605, 606 E 701

CEP: 99010051

Cidade/Estado: PASSO FUNDO

Telefone: 51-2108.6689

url: https://compass.uol e-mail: bolsas@compass.com.br

Área na empresa onde foi realizado o estágio: TI

Data de início: 02/08/2023

Data de término: 29/12/2023

Carga Horária Semanal: 20

Carga Horária Total: 400

Supervisor de Estágio: Rafael Pereira Van der Laan

APRESENTAÇÃO DA EMPRESA

A Compass UOL, empresa brasileira integrante do Grupo UOL, destaca-se como líder em serviços de transformação digital. Nossa abordagem é baseada no cultivo do talento das pessoas e na aplicação de tecnologias de ponta, como Desenvolvimento Ágil, *Multicloud*, *Data & Analytics*, Segurança, Inteligência Artificial/*Machine Learning*, APIs/Microserviços e IoT. A Compasso UOL destaca-se pelos seguintes aspectos:

• Great Place to Work, certificada pela 4ª vez consecutiva em 2022;

• AWS Partner Summit, AWS Service Partner of the Year & AWS SI Partner of the Year LATAM, premiações recebidas em 2022;

• Maior time de especialistas em Oracle Commerce do mundo;

• 100% dos clientes usando tecnologias inovadoras para transformar seus negócios;

• Mais de 10 anos de experiência em construção e gestão de nuvem e mais de 100 clientes usando Nuvem Pública através de parceiros da Compass UOL;

• 45 milhões de pedidos transacionados por ano nas plataformas desenvolvidas para os clientes;

• Clientes nos principais setores da indústria: Pagamentos e Finanças, Varejo Online e Marketplaces, Tecnologia, Mídia e Telecomunicações, Bens de Consumo, Saúde e Farmacêutica, Serviços e Outros;

• Responsável pelos principais *marketplaces* de coalizão do Brasil, atendimento das 3 maiores cias aéreas e 3 das 5 principais redes de drogarias, além das 12 empresas mais importantes do setor financeiro no mercado nacional;

• Mais de 600 mil horas em projetos B2B, B2C e D2C entregues de e-commerce em 2022;

• Mais de 2.900 colaboradores (apenas Compass UOL);

• 29 delivery centers espalhados pelo Brasil;

• Mais de 1.300 bolsas ofertadas pelo Compass Academy em 2022 e mais de 4.500 certificados e acreditações desde 2021.

A Compass UOL é especialista em projetar e construir plataformas nativas digitais para empresas líderes globais, impulsionando a inovação, transformação de negócios e prosperidade nos setores em que atuamos. A dedicação ao cultivo de talentos, criação de oportunidades e foco em tecnologias disruptivas refletem seu compromisso em impactar positivamente a sociedade.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS

API	Application Programming Interface (Interface de Programação de Aplicações)
AWS	Amazon Web Services
BI	Business Intelligence (Inteligência de Negócios)
CSV	Comma Separated Values (Valores Separados Por Vírgula)
ETL	Extract, Transform, Load (Extrair, Transformar, Carregar)
IoT	Internet of Things (Internet das Coisas)
JSON	JavaScript Object Notation (Notação de Objetos JavaScript)
SPICE	Super-fast, Parallel, In-memory Calculation Engine (Mecanismo de Cálculo Super Rápido, Paralelo e na Memória)
SQL	Structured Query Language (Linguagem de Consulta Estruturada)
TMDB	The Movie Database

SUMÁRIO

1. INTRODUÇÃO

Este relatório encapsula minha jornada enriquecedora como estagiário na Compass UOL, desempenhando o papel de Engenheiro de Dados ao longo de um período abrangente de 5 meses. Durante esse tempo, mergulhei profundamente nos fundamentos cruciais da Engenharia de Dados, com um foco especialmente afiado no processo de ETL (Extração, Transformação e Carga) de dados, utilizando os serviços robustos da AWS.

Nos estágios iniciais (Sprints 1 a 4), dediquei-me a consolidar meu conhecimento em Git, SQL, Python e Docker, estabelecendo uma base sólida para compreender os pilares essenciais da Engenharia de Dados. Cabe destacar uma "*Sprint* 0", um período de imersão nas metodologias ágeis e na segurança de aplicações Web, que se estendeu ao longo das Sprints iniciais, contribuindo significativamente para minha formação.

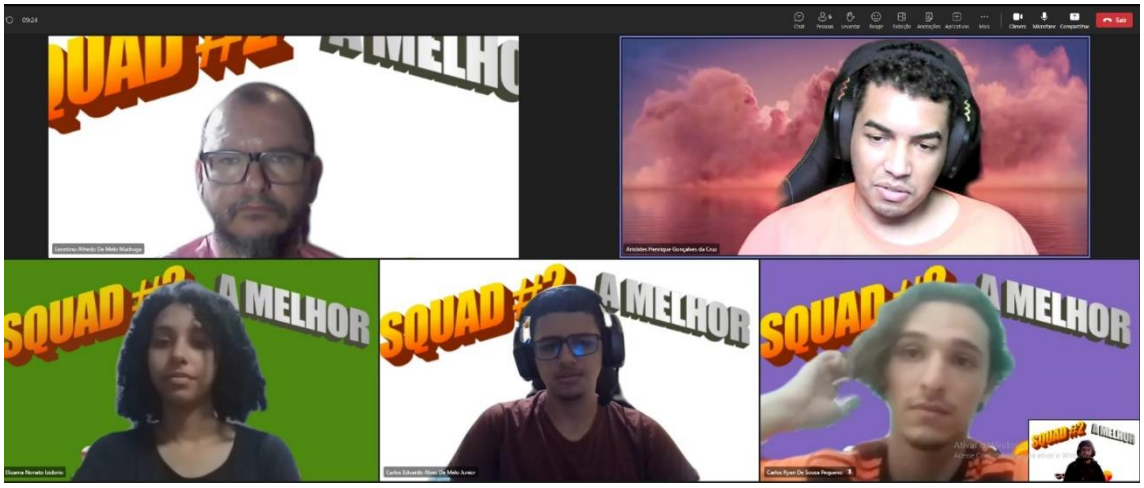
A transição para as Sprints 5 e 6 marcou minha entrada nos serviços em nuvem da AWS, expandindo consideravelmente meu escopo e preparando-me para o desafio final.

As Sprints 7 a 10 foram particularmente desafiadoras, constituindo um projeto prático onde apliquei os conhecimentos adquiridos. Este desafio envolveu a utilização das ferramentas da AWS para resolver problemas reais, centrados no vasto universo de filmes e séries, proporcionando uma análise significativa e consolidando meu aprendizado de maneira tangível.

Ao encerrar o estágio, fui agraciado com um voucher para a prova da AWS Certified Cloud Practitioner, um testemunho do compromisso da Compass UOL com o desenvolvimento profissional de seus estagiários.

Destaco, ainda, que a estrutura do estágio foi meticulosamente construída sobre os princípios organizacionais ágeis. Cada Sprint de 14 dias, liderada por um instrutor dedicado, era marcada por cerimônias de abertura e fechamento, bem como reuniões diárias, conhecidas como "dailies", garantindo um alinhamento contínuo de objetivos. E também fiz parte de uma equipe de 5 pessoas que eram denominados “Squads”, onde vivenciamos um ambiente colaborativo e de muita aprendizagem.

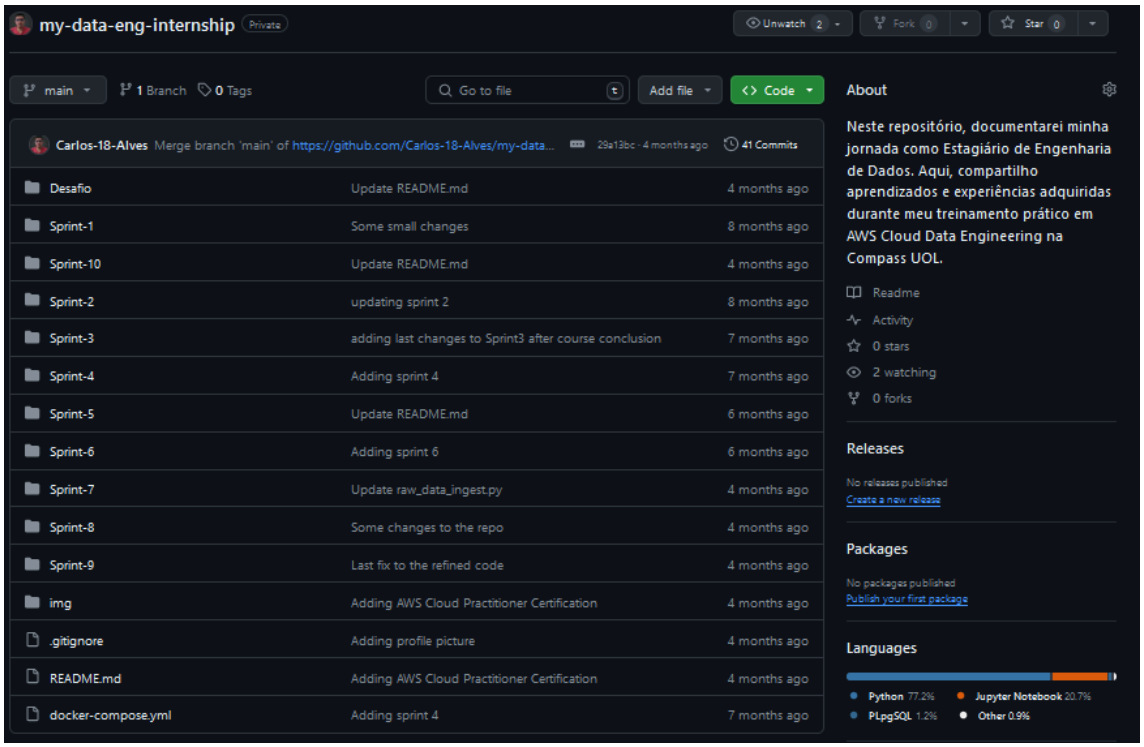
Figura 1 - Reunião do Microsoft Teams Durante a Sprint 4



Fonte: Própria Autoria (2023).

Para documentar essa jornada, criei um repositório abrangente, detalhando minha evolução como Estagiário de Engenharia de Dados, tornando transparentes meus aprendizados e experiências adquiridas durante o treinamento prático.

Figura 2 - Repositório do GitHub onde documentei o estágio



Fonte: Própria Autoria (2024).

Esta introdução visa proporcionar uma visão holística do meu estágio, delineando o percurso desde a consolidação dos fundamentos até a aplicação prática, culminando na oferta de certificação ao término do programa. Nos próximos capítulos, detalharei minhas atividades, conquistas e reflexões, oferecendo uma fundação sólida para a avaliação do meu desempenho no estágio curricular não obrigatório supervisionado no IFPB.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Este trabalho tem como objetivo geral relatar as atividades desenvolvidas na Compass UOL, durante o período em que atuei como estagiário em Engenharia de Dados.

2.2. OBJETIVOS ESPECÍFICOS

a) Descrever as atividades desempenhadas durante o estágio em Engenharia de Dados, incluindo a resolução de desafios e a implementação de soluções;
b) Descrever as ferramentas e tecnologias utilizadas;
c) Apresentar uma experiência profissional na Compass UOL, destacando aprendizados, desafios enfrentados e perspectivas sobre o mercado de Engenharia de Dados.

3. TECNOLOGIAS E FERRAMENTAS

Nesta seção, descreverei as tecnologias e ferramentas que utilizei ao longo da minha experiência de estágio, fundamentais para compreender as atividades desenvolvidas por mim.

3.1. GIT E GITHUB

O Git é um sistema de controle de versão distribuído amplamente utilizado no desenvolvimento de software. Ele permite que os desenvolvedores colaborem em projetos, acompanhem as alterações no código-fonte ao longo do tempo e revertam para versões anteriores, se necessário. O Git é especialmente útil em ambientes de desenvolvimento colaborativo, pois facilita o trabalho em equipe, mantém um histórico detalhado das alterações e ajuda a evitar conflitos de código (GUEDES, 2019).

O GitHub, por sua vez, é uma plataforma de hospedagem de código-fonte baseada na web que utiliza o Git para controle de versão. Ele oferece recursos adicionais para colaboração em projetos, como rastreamento de problemas, gerenciamento de tarefas, revisão de código e integração contínua. O GitHub é amplamente utilizado pela comunidade de desenvolvimento de software para compartilhar, colaborar e contribuir para uma variedade de projetos de código aberto e privados (GUEDES, 2019).

Juntos, o Git e o GitHub formam uma poderosa combinação para o desenvolvimento de software colaborativo, permitindo que os desenvolvedores trabalhem de forma eficiente em equipe, gerenciem e acompanhem as alterações no código-fonte e colaborem em projetos de maneira transparente.

3.2. SQL

SQL (*Structured Query Language)* é uma linguagem de programação utilizada para gerenciar e manipular bancos de dados relacionais.

Com o SQL, os desenvolvedores podem executar uma variedade de operações em bancos de dados, como recuperar dados específicos, inserir novos dados, atualizar registros existentes e excluir informações. A linguagem

oferece uma sintaxe simples e poderosa para expressar consultas e comandos de manipulação de dados, permitindo que os usuários realizem uma ampla gama de tarefas de gerenciamento de banco de dados (AWS, 2023).

3.3. PYTHON

Python é uma linguagem de programação versátil, amplamente utilizada em diversas áreas. Sua popularidade decorre de sua eficiência, facilidade de aprendizado e capacidade de execução em várias plataformas. Com uma vasta biblioteca padrão e uma comunidade ativa, Python permite aos desenvolvedores realizar tarefas complexas com menos código, facilitando a produtividade e a colaboração. Além disso, é amplamente empregado em aplicações como desenvolvimento web, automação de *scripts*, ciência de dados, *machine learning* e desenvolvimento de software, oferecendo uma solução abrangente para diversas necessidades de programação (AWS, 2023).

3.4. DOCKER

O Docker é uma plataforma de virtualização que utiliza o conceito de containers para facilitar a distribuição e execução de software de forma isolada e eficiente. Ao contrário das virtualizações convencionais, que exigem a instalação de um software específico para gerenciar máquinas virtuais, o Docker aproveita os recursos do sistema operacional subjacente, como o *kernel* do Linux, para criar ambientes isolados conhecidos como containers (Hostinger, 2024).

Esses containers encapsulam o software e todas as suas dependências, garantindo que ele seja executado de maneira consistente em diferentes ambientes, desde o desenvolvimento até a produção. Com o Docker, é possível transportar aplicativos de um ambiente para outro de forma rápida e segura, sem se preocupar com a configuração do sistema operacional ou outros pré-requisitos (Hostinger, 2024).

3.5. AWS

A AWS é uma plataforma de computação em nuvem líder mundial, oferecendo uma ampla gama de serviços distribuídos em data centers ao redor do globo. Com mais de 200 serviços disponíveis, a AWS atende a uma variedade de necessidades, desde infraestrutura básica, como computação e armazenamento, até tecnologias emergentes, como machine learning, inteligência artificial, análise de dados e IoT (Internet das Coisas) (AWS, 2023).

Um dos principais diferenciais da AWS é sua funcionalidade máxima. A plataforma oferece uma variedade maior de serviços e recursos do que qualquer outro provedor de nuvem, facilitando e agilizando a migração de aplicativos para a nuvem e o desenvolvimento de soluções inovadoras. Além disso, a AWS se destaca pela profundidade de suas ofertas, fornecendo uma ampla gama de opções, como diferentes tipos de bancos de dados, para atender às necessidades específicas de cada cliente (AWS, 2023).

A seguir irei descrever alguns dos principais serviços AWS que utilizei durante o meu estágio:

Amazon S3: O Amazon S3 (Amazon Simple Storage Service) é um serviço de armazenamento de objetos altamente escalável, seguro e de alto desempenho oferecido pela AWS. Ele é usado para armazenar uma ampla variedade de dados, como dados de produção, backups, arquivos de mídia, entre outros, para diversos casos de uso, como *data lakes*, *websites*, aplicativos móveis, e análises de big data (AWS, 2024).

AWS Lambda: O AWS Lambda é um serviço de computação oferecido pela AWS que permite executar código sem a necessidade de provisionar ou gerenciar servidores. Ele opera seu código em uma infraestrutura altamente disponível, realizando toda a administração dos recursos computacionais, incluindo manutenção do servidor, provisionamento, escalabilidade automática e registro de logs (AWS, 2024).

AWS Glue: O AWS Glue é um serviço sem servidor de integração de dados que simplifica a descoberta, preparação, transferência e integração de dados de várias fontes para análise, machine learning e desenvolvimento de aplicações. Com suporte para mais de 70 fontes de dados, o Glue oferece uma plataforma para criar, executar e monitorar visualmente pipelines de ETL,

gerenciando dados em um catálogo centralizado. Ele integra-se facilmente aos serviços de análise da AWS e aos data lakes do Amazon S3, proporcionando escalabilidade sob demanda e um modelo de pagamento conforme o uso (AWS, 2024).

Amazon QuickSight: O Amazon QuickSight é um serviço de inteligência de negócios (BI) em escala de nuvem que oferece insights compreensíveis de maneira fácil para colaboradores, conectando dados de diversas fontes em um único painel. Com recursos como o mecanismo na memória SPICE (Mecanismo de Cálculo Super Rápido, Paralelo e na Memória), análise colaborativa e visualização interativa, a QuickSight permite a exploração e interpretação de dados de qualquer dispositivo, proporcionando benefícios como velocidade, baixo custo total de propriedade, análises automatizadas por machine learning, segurança empresarial e *pay-per-session pricing*. Para começar, os usuários podem explorar recursos como terminologia essencial, análise de dados, e segurança (AWS, 2024).

3.6. TMDB

O TMDB (The Movie Database) é uma base de dados gratuita e de código aberto sobre filmes e séries, criada em 2008 por Travis Bell. Inicialmente focada em filmes, em 2013 foi expandida para incluir também séries. O TMDB é alimentado pela contribuição da comunidade e é uma fonte crucial de dados para vários serviços. Com mais de 300 mil filmes e crescendo, o TMDB processa diariamente milhares de edições feitas pelos usuários, moderadas por voluntários. Seu acesso é gratuito para entidades não comerciais, e sua API (Interface de Programação de Aplicações) é amplamente utilizada para integração de dados em vários idiomas (Wikipédia, 2022).

4. ATIVIDADES DESENVOLVIDAS

No decorrer das sprints 1 a 10, embarquei em uma jornada de aprendizado enriquecedora como estagiário na Compasso UOL, atuando na função de estagiário em Engenharia de Dados. Esta seção oferece uma visão panorâmica das atividades desenvolvidas, desde a consolidação dos fundamentos até desafios práticos e a aplicação de conhecimentos avançados em Engenharia de Dados.

4.1. SPRINTS INICIAIS: DESENVOLVIMENTO DE CONHECIMENTOS BÁSICOS

Esta subseção detalha as atividades realizadas nas quatro primeiras sprints, focando na consolidação de conhecimentos fundamentais em Git, SQL, Python e Docker. Além disso, explora-se o aprendizado de conceitos essenciais de segurança em aplicações web e metodologias ágeis.

4.1.1. O que foi feito?

Durante essas sprints, dediquei-me à consolidação de conhecimentos fundamentais em Git, SQL, Python e Docker. Essa fase não apenas estabeleceu as bases essenciais para as atividades subsequentes, incluindo o desafio final, mas também abrangeu o aprendizado de conceitos cruciais de segurança em aplicações web e metodologias ágeis.

4.1.2. Por que foi feito?

Essa fase inicial foi um pilar fundamental para construir uma base sólida, capacitando-me para desafios mais avançados no campo da Engenharia de Dados. A compreensão aprofundada dessas ferramentas e conceitos essenciais foi vital para enfrentar tarefas mais complexas posteriormente.

4.1.3. Como foi feito?

Engajei-me em cursos especializados na plataforma Udemy, onde não apenas absorvi conhecimentos teóricos, mas também apliquei ativamente esses conceitos em atividades práticas. Fazer anotações detalhadas e participar ativamente de projetos práticos foi parte integrante deste processo de aprendizado.

2.1.4. Qual a aprendizagem com a atividade?

Esta etapa inicial foi essencial para aprimorar minha proficiência em Git, Python e SQL. A compreensão profunda do funcionamento do Docker foi especialmente destacada, pois reconheci sua importância como uma ferramenta-chave na criação e gestão eficiente de ambientes de desenvolvimento. Além disso, o aprendizado abrangente de segurança em aplicações web e metodologias ágeis contribuiu para a minha formação holística como um profissional de Engenharia de Dados, preparando-me para os desafios subsequentes com confiança e competência.

4.2. Introdução aos Serviços em Nuvem da AWS

Esta seção destaca as sprints 5 e 6, que marcaram minha imersão nos serviços em nuvem da AWS. Durante esse período, meu foco estava na compreensão abrangente e no aprendizado prático para iniciar o desafio final.

4.2.1. O que foi feito?

Durante as sprints 5 e 6, dediquei-me a uma exploração aprofundada dos serviços em nuvem oferecidos pela AWS. Essa fase concentrou-se na compreensão abrangente e no aprendizado prático desses serviços inovadores.

4.2.2. Por que foi feito?

O objetivo principal era assimilar os conceitos fundamentais dos serviços em nuvem da AWS, aprimorando minha familiaridade com essas ferramentas essenciais na área de engenharia de dados. Essa iniciativa visava alinhar-me às melhores práticas e tendências emergentes nesse campo dinâmico e também me auxilia para quando fosse iniciar o desafio final.

4.2.3. Como foi feito?

Envolveu minha participação ativa em treinamentos especializados disponíveis na plataforma AWS, abordando tópicos específicos relacionados aos serviços em nuvem. Além disso, aplicação prática dos conhecimentos adquiridos em atividades de fixação, garantindo uma sólida compreensão e habilidade prática.

4.2.4. Qual a aprendizagem com a atividade?

Este período resultou na aquisição de uma expertise substancial em ambientes em nuvem, destacando não apenas a compreensão teórica, mas também a aplicação prática efetiva desses serviços. Adquiri uma compreensão holística da implementação prática, reconhecendo a importância estratégica desses serviços e discernindo as vantagens práticas de sua utilização no contexto da engenharia de dados.

4.3. Desafio Etapa I: Ingestão de Arquivos CSV no Amazon S3

Nesta etapa, o desafio consistiu na ingestão de arquivos CSV (Valores Separados Por Vírgula) no Amazon S3, uma parte crucial do processo de construção do *data lake*.

4.3.1. O que foi feito?

Durante essa atividade, desenvolvi um script em Python para acessar o armazenamento local, extrair os dados de filmes e séries que nos foram disponibilizados e realizar a ingestão desses dados no *Bucket* do Amazon S3, que é um sistema de armazenamento em nuvem. Utilizei um contêiner Docker para executar esse script, efetuando assim a transferência dos dados e criando nossa "Zona *Raw*" para armazenar os dados brutos.

4.3.2. Por que foi feito?

A separação dos dados por camadas é uma prática essencial para um engenheiro de dados. A criação da camada *Raw* é o ponto inicial, onde dados brutos são trazidos e armazenados em uma área específica do nosso *data lake*, aguardando processamentos posteriores.

4.3.3. Como foi feito?

O processo envolveu a criação de um script Python que:

- Leu os dados de filmes e séries a partir do armazenamento local.
- Utilizou a biblioteca boto3 para interagir com os serviços do Amazon S3.
- Definiu a estrutura de diretórios no S3, incluindo a camada *Raw* e pastas organizadas por ano, mês e dia.
- Carregou os arquivos CSV no S3, seguindo a estrutura de diretórios definida.

Figura 3 - Trecho do código Python para carregamento de dados no Amazon S3

```python
import boto3
import os
from datetime import datetime

# Configuração de credenciais da AWS e região
AWS_ACCESS_KEY_ID = os.environ.get('AWS_ACCESS_KEY_ID')
AWS_SECRET_ACCESS_KEY = os.environ.get('AWS_SECRET_ACCESS_KEY')
AWS_REGION = 'us-east-1'

TMDB_API_TOKEN = os.environ.get('TMDB_API_TOKEN')
# Definindo o nome do bucket S3 e caminhos
NOME_DO_BUCKET_S3 = 'data-lake-do-carlos-alves'
ZONA_RAW = 'Raw/Local/CSV'
data_atual = datetime.now().strftime('%Y/%m/%d')

# Definindo os caminhos locais dos arquivos CSV (essa pasta foi colocada no git ignore por serem arquivos grandes)
caminho_local_filmes_csv = 'files/movies.csv'
caminho_local_series_csv = 'files/series.csv'

# Inicializando o cliente S3
s3 = boto3.client('s3',
    aws_access_key_id=AWS_ACCESS_KEY_ID,
    aws_secret_access_key=AWS_SECRET_ACCESS_KEY,
    region_name=AWS_REGION
)

# Defininido os caminhos no S3 para os arquivos
chave_s3_filmes = f"{ZONA_RAW}/Movies/{data_atual}/{os.path.basename(caminho_local_filmes_csv)}"
chave_s3_series = f"{ZONA_RAW}/Series/{data_atual}/{os.path.basename(caminho_local_series_csv)}"

# Fazendo o upload dos arquivos CSV para o S3
s3.upload_file(caminho_local_filmes_csv, NOME_DO_BUCKET_S3, chave_s3_filmes)
s3.upload_file(caminho_local_series_csv, NOME_DO_BUCKET_S3, chave_s3_series)

print("Arquivos CSV enviados com sucesso para o S3.")

```

Fonte: Própria Autoria (2024).

4.3.4. Qual a aprendizagem com a atividade?

Esta atividade proporcionou uma compreensão mais profunda dos serviços do Amazon S3 e da importância de automatizar processos. A capacidade de transferir dados de forma remota, sem a necessidade de uploads manuais, destaca a eficiência da automação, tornando o ambiente de trabalho mais ágil e propenso a futuras expansões no *data lake*. Essa automação não apenas agiliza a ingestão de dados, mas também contribui para a consistência e confiabilidade do processo.

4.4. Desafio Etapa II: Captura de dados do TMDB

Nesta etapa, o desafio consistiu na captura de dados específicos de filmes e séries do TMDB, com foco especial em séries de crime.

4.4.1. O que foi feito?

Nesta fase, realizei chamadas de API do TMDB por meio do AWS Lambda para capturar e armazenar dados específicos de filmes e séries, focando especialmente em séries de crime.

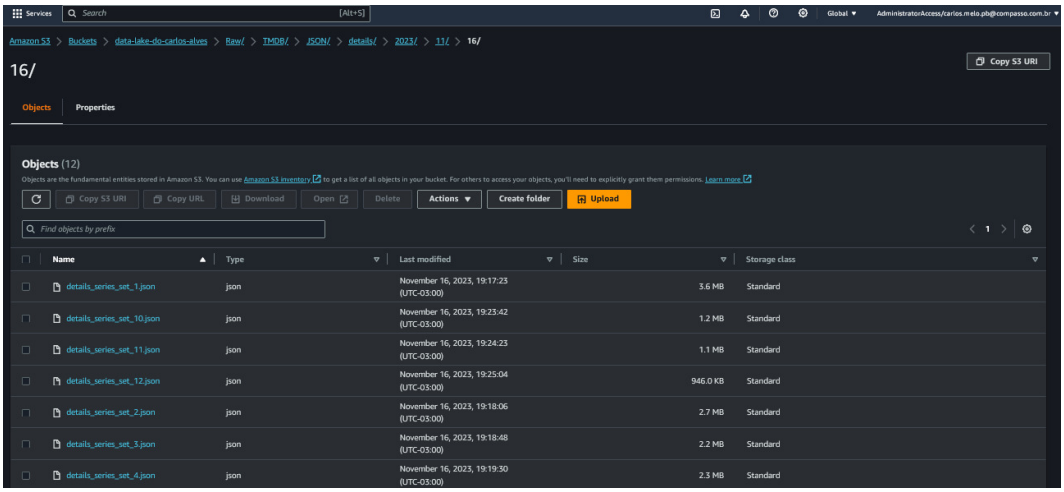
4.4.2. Por que foi feito?

A realização desse desafio tinha como propósito enriquecer ainda mais meu *data lake*. Ao incorporar informações adicionais sobre séries de crime do TMDB, pretendia aprimorar a qualidade e a amplitude dos dados disponíveis para análise.

4.4.3. Como foi feito?

Para executar essa tarefa, criei um código Lambda que filtrava especificamente as séries de crime no TMDB. Utilizando uma função Lambda integrada aos serviços da AWS, configurei uma Role que concedia acesso ao meu *data lake* no Amazon S3. O acesso aos dados da API do TMDB foi garantido através de tokens, permitindo-me recuperar informações sobre 1200 séries de crime, que ficaram separadas em arquivos com 100 cada, ao todo 12. Esses dados foram então armazenados na camada *Raw* do meu ambiente.

Figura 4 - Estrutura de diretórios no Amazon S3 após rodar o código



Fonte: Própria Autoria (2023).

4.4.4. Qual a aprendizagem com a atividade?

Essa atividade proporcionou uma compreensão valiosa sobre o uso do AWS Lambda para acessar dados de APIs externas, neste caso, do TMDB. A criação de uma Role com permissões específicas eliminou a necessidade de criar um usuário com credenciais incorporadas no código, simplificando a gestão de acesso e reforçando a segurança do processo. Essa abordagem demonstrou a flexibilidade e eficácia do AWS Lambda como uma ferramenta para integração de dados de fontes externas em ambientes de *data lake*.

4.5. Desafio Etapa III: Modelagem e Refinamento dos Dados

Neste desafio, o foco foi o processamento e refinamento dos dados no contexto do *data lake*, com o objetivo de prepará-los para análises mais aprofundadas.

4.5.1. O que foi feito?

Durante esta atividade, efetuei o processamento da camada *Trusted* e a modelagem e processamento da camada *Refined* no contexto do meu Data Lake. Na camada *Trusted*, desenvolvi scripts ETL usando AWS Glue com Apache Spark para processar dados brutos da *Raw Zone*, removendo atributos desnecessários e consolidando-os em arquivos Parquet armazenados na camada *Trusted*. Na fase de modelagem, utilizei os dados da *Trusted* para criar tabelas fato e dimensão, preparando-os para análises mais aprofundadas.

4.5.2. Por que foi feito?

Essa atividade foi crucial para aprimorar a qualidade dos dados e criar uma estrutura organizada para análises subsequentes. O processamento da camada *Trusted* garantiu que os dados brutos fossem transformados e limpos antes de serem refinados, enquanto a modelagem da camada *Refined* preparou o terreno para questões específicas a serem abordadas na próxima etapa.

4.5.3. Como foi feito?

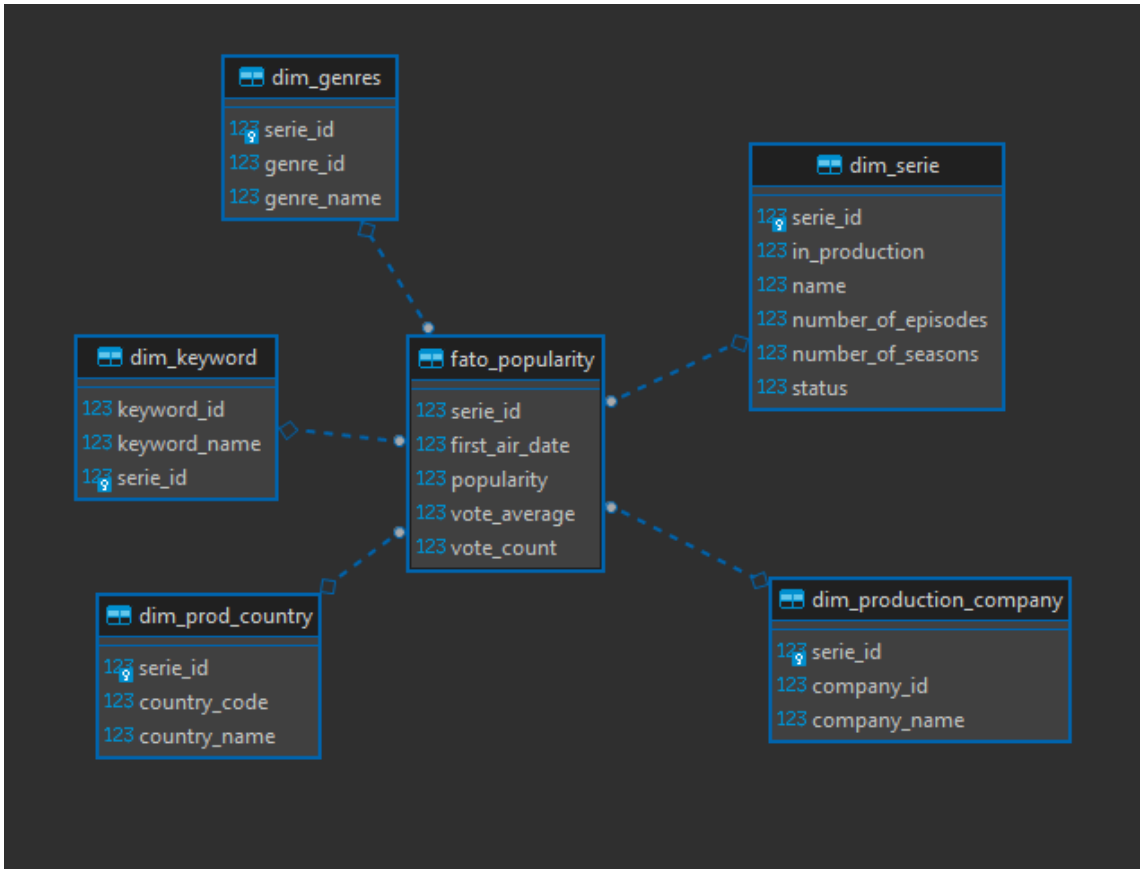
Inicialmente, examinei amostras dos arquivos JSON (Notação de Objetos JavaScript) dos dados brutos para identificar e abordar inconsistências. No tratamento dos dados, removi atributos desnecessários e tratei valores nulos, mantendo a integridade dos dados.

Desenvolvi um código no AWS Glue para processar os dados brutos, realizando transformações necessárias e salvando-os na camada *Refined*. Esse processo foi executado de forma eficiente, garantindo a consistência dos dados. Modelagem da *Refined*:

Defini as tabelas fato e dimensão para a camada *Refined* com o objetivo de responder perguntas específicas sobre séries de crime. As tabelas incluem:

- dim_serie: informações relevantes sobre as séries, como nome, número de episódios, status de produção, etc.
- dim_keyword: informações sobre palavras-chave associadas às séries, identificando preferências do público.
- dim_production_company: dados sobre as empresas produtoras das séries.
- dim_genres: gêneros das séries, incluindo além de crime, outros gêneros.
- dim_prod_country: país de produção das séries.
- fato_popularity: dados de popularidade da série, média de votos, contagem de votos.

Figura 5 - Diagrama da Modelagem da Camada *Refined*



Fonte: Própria Autoria (2023).

4.5.4. Qual a aprendizagem com a atividade?

Nesta etapa, explorei a modelagem relacional e dimensional, enfrentando desafios que se tornaram mais acessíveis com a assimilação dos conceitos fundamentais. A manipulação da Camada *Trusted* e a modelagem na Camada *Refined* não apenas proporcionaram uma base sólida para análises futuras, mas também representaram uma conquista pessoal e um alicerce robusto para o crescimento profissional no campo da Engenharia e Análise de Dados.

4.6. Desafio Etapa IV: Criação do Dashboard

Neste desafio, o foco foi na criação de um Dashboard no QuickSight, utilizando os dados refinados provenientes das etapas anteriores do projeto.

4.6.1. O que foi feito?

Durante a atividade, concentrei esforços na criação do Dashboard no QuickSight, utilizando os dados refinados provenientes das etapas anteriores do projeto. Inicialmente, participei de um curso abrangente sobre o QuickSight para aprimorar meu entendimento das principais funcionalidades da ferramenta. Essa capacitação foi crucial para a elaboração eficiente e otimizada do Dashboard.

4.6.2. Por que foi feito?

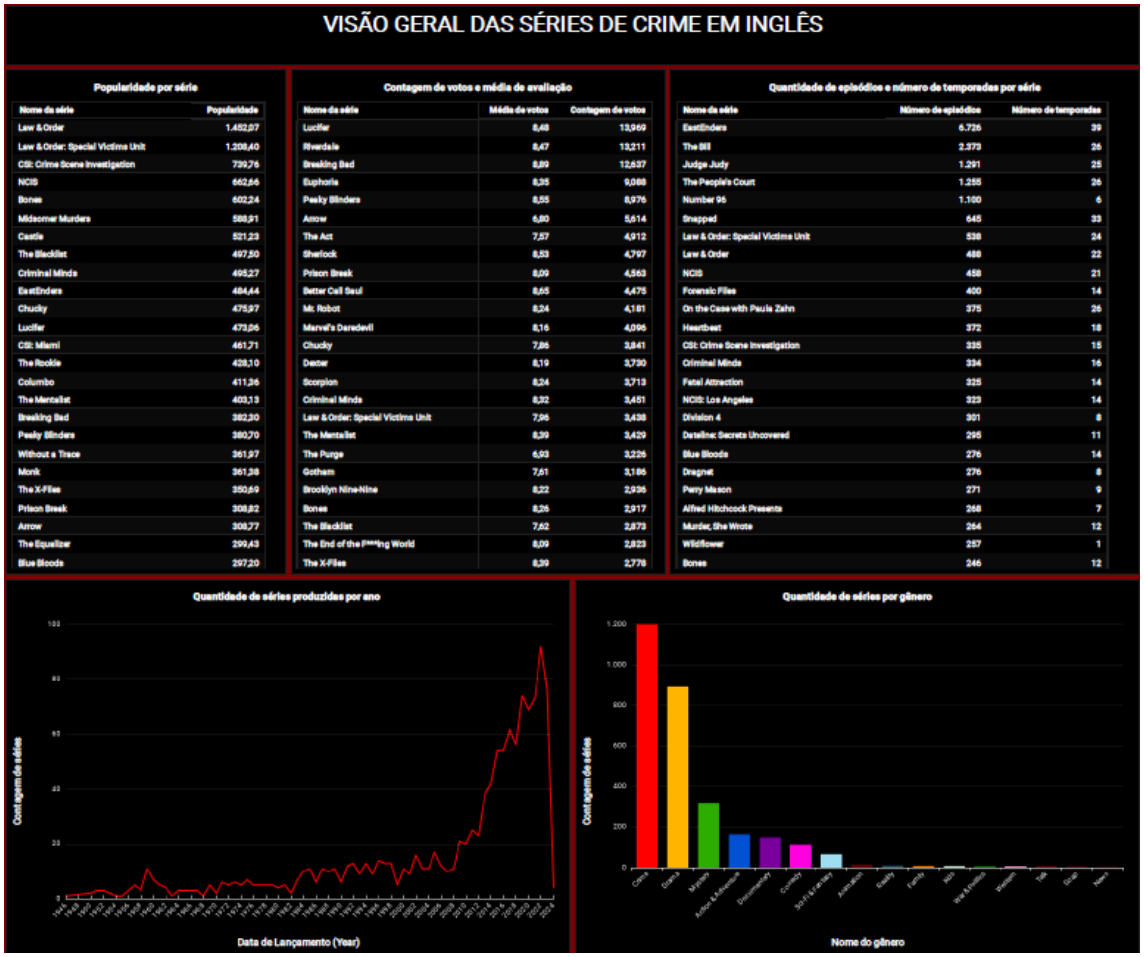
A criação do Dashboard visa proporcionar uma visão acessível e interativa das análises realizadas no projeto. Essa ferramenta permite a visualização rápida e clara de insights, facilitando a tomada de decisões informadas.

4.6.3. Como foi feito?

Decidi estruturar o Dashboard em três *sheets* (ou páginas), cada uma abordando diferentes aspectos dos dados e da modelagem dimensional realizada:

Visão Geral: possui informações abrangentes sobre as séries, destacando as mais populares, bem avaliadas, quantidade de episódios, distribuição por época, gênero e detalhes sobre as empresas produtoras.

Figura 6 - Parte da *sheet* Visão Geral do QuickSight



Fonte: Própria Autoria (2023).

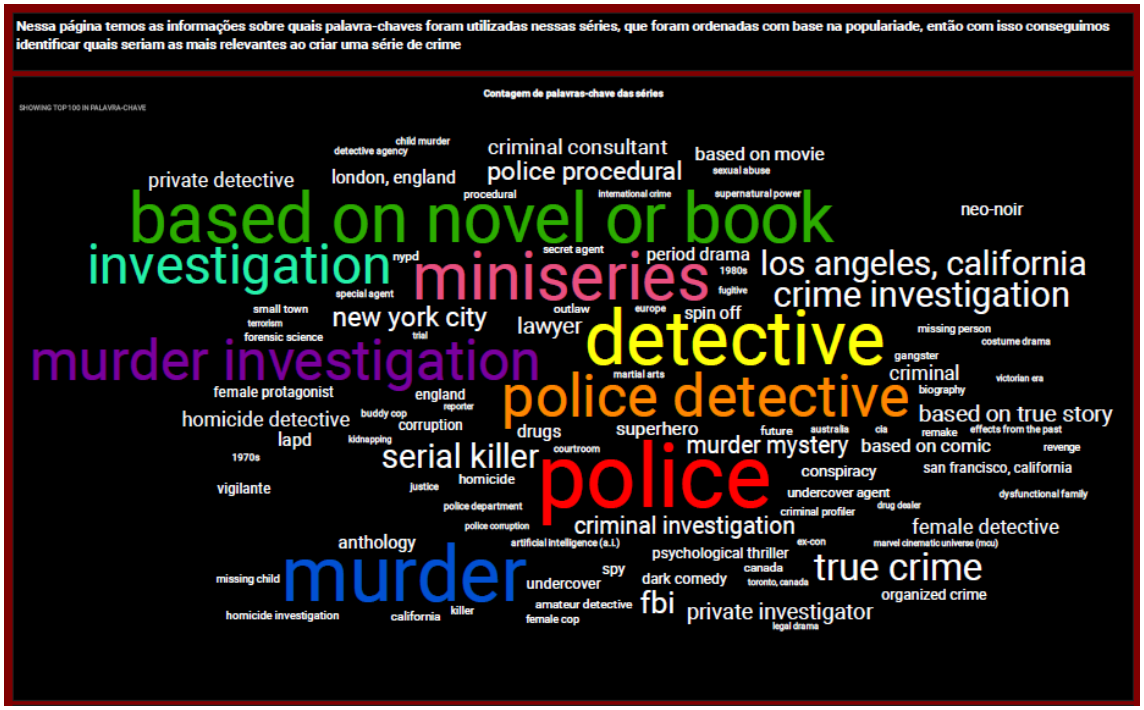
Comparação entre Séculos: foquei na comparação entre séries do Século XX e XXI, destacando diferenças e tendências ao longo do tempo.

Figura 7 - Parte da sheet de Comparação entre séculos no QuickSight

Fonte: Própria Autoria (2023).

Palavras-chave: destaque para as palavras-chave mais utilizadas em séries de crime, fornecendo insights valiosos para a criação de novas séries.

Figura 8 - Parte da sheet de palavras-chave no QuickSight



Fonte: Própria Autoria (2023).

4.6.4. Qual a aprendizagem com a atividade?

A criação do Dashboard representou uma integração eficaz de conhecimentos adquiridos ao longo do projeto. A participação no curso do QuickSight fortaleceu minhas habilidades na ferramenta, e a aplicação prática na construção do Dashboard solidificou a compreensão dos conceitos. Essa etapa não apenas demonstrou a capacidade de visualizar análises complexas de maneira simplificada, mas também ressaltou a importância de uma apresentação clara e intuitiva para comunicar efetivamente *insights* a diferentes públicos.

5. CONSIDERAÇÕES FINAIS

Ao finalizar este relatório, é gratificante refletir sobre a jornada realizada durante este projeto de engenharia e análise de dados. As múltiplas etapas, desafios e aprendizados ao longo das sprints contribuíram para uma compreensão mais profunda e habilidosa do universo da manipulação, processamento e análise de dados.

Iniciamos com a definição do escopo e dos desafios a serem enfrentados, passando pela coleta de dados brutos até a criação de um Dashboard interativo. Cada etapa foi cuidadosamente planejada, executada e refinada, demonstrando a importância da metodologia ágil e da flexibilidade diante das complexidades do processo.

A abordagem de camadas no Data Lake revelou-se crucial para a organização eficiente dos dados, desde a Zona Bruta até a camada *Refined*, onde os dados foram modelados para análises específicas. A utilização de serviços da AWS, como S3, Lambda e Glue, proporcionou uma infraestrutura robusta e escalável para suportar o processamento e a análise de grandes volumes de dados.

A modelagem relacional e dimensional, embora desafiadora, evidenciou sua importância na estruturação dos dados para análises mais avançadas. A compreensão da relação entre as tabelas fato e dimensão permitiu a formulação de perguntas analíticas mais precisas, resultando em insights mais relevantes.

A criação do Dashboard no QuickSight trouxe os dados à vida, transformando números e estatísticas em visualizações intuitivas e impactantes. A capacidade de comunicar efetivamente os insights obtidos é tão vital quanto a própria análise, e o Dashboard se destacou como uma ferramenta essencial para essa finalidade.

Além das habilidades técnicas, a importância da colaboração, comunicação e adaptação durante o processo tornou-se evidente. A interação contínua com a equipe, a resolução de desafios inesperados e a celebração das conquistas refletem não apenas o sucesso técnico, mas também a maturidade e coesão do time.

Em última análise, este projeto não é apenas um registro de atividades técnicas, mas uma narrativa de superação, aprendizado contínuo e realização

profissional. Os conhecimentos adquiridos neste percurso representam não apenas habilidades técnicas, mas uma base sólida para enfrentar os desafios crescentes no cenário dinâmico da engenharia de dados.

Este relatório, portanto, não marca apenas o fim de um projeto, mas o início de uma trajetória contínua de crescimento e excelência na engenharia e análise de dados.

REFERÊNCIAS

GUEDES, Marylene. **Git e GitHub: quais as diferenças?**. TreinaWeb, 2019. Disponível em: https://www.treinaweb.com.br/blog/git-e-github-quais-as-diferencas. Acesso em 13 de abr. de 2024.

O que é SQL (linguagem de consulta estruturada)?. AWS, 2023. Disponível em: https://aws.amazon.com/pt/what-is/sql/. Acesso em 13 de abr. de 2024.

O que é Python?. AWS, 2023. Disponível em: https://aws.amazon.com/pt/what-is/python/. Acesso em 13 de abr. de 2024.

O Que é Docker e Como Ele Funciona? – Docker Explicado. Hostinger, 2024. Disponível em: https://www.hostinger.com.br/tutoriais/o-que-e-docker. Acesso em 13 de abr. de 2024.

O que é AWS? Como funciona Amazon Web Services Amazon. AWS, 2023. Disponível em: https://aws.amazon.com/pt/what-is-aws/. Acesso em 13 de abr. de 2024.

O que é o Amazon S3?. AWS, 2024. Disponível em: https://docs.aws.amazon.com/pt_br/AmazonS3/latest/userguide/Welcome.html. Acesso em 13 de abr. de 2024.

O que é o AWS Lambda?. AWS, 2024. Disponível em: https://docs.aws.amazon.com/pt_br/lambda/latest/dg/welcome.html. Acesso em 13 de abr. de 2024.

O que é AWS Glue?. AWS, 2024. Disponível em: https://docs.aws.amazon.com/pt_br/glue/latest/dg/what-is-glue.html. Acesso em 13 de abr. de 2024.

O que é a Amazon QuickSight?. AWS, 2024. Disponível em: https://docs.aws.amazon.com/pt_br/quicksight/latest/user/welcome.html. Acesso em 13 de abr. de 2024.

THE MOVIE DATABASE. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2022. Disponível em: https://pt.wikipedia.org/w/index.php?title=The_Movie_Database&oldid=63560618. Acesso em: 14 de abr. de 2024.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Monteiro - Código INEP: 25284940
Pb-264, S/N, Serrote, CEP 58500-000, Monteiro (PB)
CNPJ: 10.783.898/0008-41 - Telefone: (83) 3351-3700

Documento Digitalizado Ostensivo (Público)

Relatório de Estágio

Assunto:	Relatório de Estágio
Assinado por:	Carlos Alves
Tipo do Documento:	Relatório
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Carlos Eduardo Alves de Melo Júnior**, **ALUNO (202015020004) DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - MONTEIRO**, em 10/09/2024 16:03:54.

Este documento foi armazenado no SUAP em 10/09/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1244741
Código de Autenticação: ce2556c7d0