**Documentação de Arquitetura de Dados**

1. Introdução

Esta documentação tem como propósito proporcionar uma compreensão clara da arquitetura dos dados, fornecendo recomendações práticas para aprimorar a confiabilidade e qualidade dos dados ao longo do tempo. As orientações visam enriquecer a documentação, garantindo abrangência, acessibilidade e eficácia para orientar usuários na interação e compreensão da arquitetura de dados, incluindo práticas recomendadas.

2. Estrutura do Banco de Dados MySQL:

- Diagrama ER (Entidade-Relacionamento) para a visualização dos relacionamentos entre as tabelas:

Interface gráfica do usuário

Descrição gerada automaticamente

Figura - Relacionamento entre tabelas (MySQL).

O Diagrama ER, que significa Entidade-Relacionamento, é uma representação gráfica que descreve as entidades (objetos ou conceitos) dentro de um sistema e os relacionamentos entre essas entidades. É comumente usado em bancos de dados para visualizar e projetar a estrutura de um banco de dados.

Ao criar um Diagrama ER para um banco de dados MySQL, pode-se visualizar de forma clara e concisa como as tabelas estão relacionadas entre si. Isso ajuda na compreensão da estrutura do banco de dados e na identificação de áreas onde podem ser necessárias modificações ou otimizações. O Diagrama ER facilita a comunicação entre desenvolvedores, designers de banco de dados e outros stakeholders envolvidos no projeto.

- Exemplos de consultas SQL frequentemente utilizadas para melhor compreensão:

Texto

Descrição gerada automaticamente

Figura - Parte de código que descreve rotinas utilizadas em MySQL.

Realizamos as conexões entre as tabelas e a atribuição das chaves, bem como a definição dos tipos de dados (Text, Varchar, etc.), por meio de linhas de código SQL. Esse processo garante a integridade e eficiência da estrutura do banco de dados, permitindo a gestão precisa das relações entre as diferentes entidades. Através do SQL, conseguimos não apenas visualizar, mas também implementar de maneira prática e eficaz os relacionamentos e características fundamentais de nossas tabelas no MySQL.

Texto

Descrição gerada automaticamente

Figura - Parte de código que modifica variáveis em uma tabela do MySQL.

Texto

Descrição gerada automaticamente

Figura - Parte de código que insere uma chave estrangeira, criando o relacionamento entre tabelas no MySQL.

3. Modelos de Dados no Power BI:

As imagens a seguir representam medidas importantes realizadas com o auxílio da ferramenta do Power BI.

- Soma do Custo de Total de Entrega x Ano utilizando tanto modelos de aviação comercial quanto os modelos de aviação executiva.

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Figura - Comparação da Soma do Custo Total de Frete ao longo dos anos na Aviação Comercial.

Uma imagem contendo Aplicativo

Descrição gerada automaticamente

Figura - Comparação da Soma do Custo Total de Frete ao longo dos anos na Aviação Executiva.

- Média do Custo do Combustível (QAV) por período, sendo o período caracterizado por pré-pandemia (antes de 2020) e pós-pandemia (2020 em diante), e o impacto no frete para os mesmos períodos

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

Figura - A Figura a direita representa a média do custo do combustível (QAV) por período (pré e pós pandemia). A Figura a esquerda representa a média do custo de frete por período.

- A quantidade de aeronaves entregue por ano e por modelo, tanto para o segmento de Aviação Comercial (Figura 8), quanto para o segmento de Aviação Executiva (Figura 9).

Gráfico, Histograma

Descrição gerada automaticamente

Figura - A primeira análise (acima) representa a quantidade de aeronaves entregue pela Embraer por ano. A Segunda análise (abaixo) representa a quantidade de aeronaes entregue por modelo na Aviação Comercial.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Figura - A primeira análise (acima) representa a quantidade de aeronaves entregue pela Embraer por ano. A Segunda análise (abaixo) representa a quantidade de aeronaes entregue por modelo na Aviação Executiva.

- Uma análise que representa a eficiência de entregas da Embraer relacionadas a Aviação Comercial para os anos de 2003 a 2023 (85,8% dos pedidos realizados já foram entregue).

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente com confiança média

Figura - Gráfico de pizza que mostra a quantidade de aeronaves entregue por pedidos firmes em carteira do segmento de Aviação Comercial.

- Lógica das Transformações:

As transformações realizadas foram limpeza, normalização e aumento de dados.

A limpeza inclui retirada de informações não usais (como, por exemplo, ‘Clientes não-divulgados’) que foram substituídas por valores gerados aleatoriamente entre os valores já existentes. Além da padronização dos valores caracterizados por *string*.

A normalização entre o preço de custo de frete de 1 (um) avião para a quantidade de aviões entregue para um cliente.

Custo de frete: ,  
onde tamanho do tanque de combustível da aeronave, valor do combustível para determinado ano, Autonomia da aeronave e Distância ao país de destino a partir de São José dos Campos.

Custo total de frete: ,

onde é a quantidade de aeronaves entregue a determinado país.

O aumento de dados foi realizado com base na quantidade de aviões entregue relacionado aos dados faltantes de cada tabela. Se a quantidade de aviões entregue é maior do , dividimos em valores inteiros aleatórios com a condição de que a soma seja igual a . Por exemplo, , geramos aleatoriamente e tal que , ou , e , tal que, nestes casos seria igual a e igual a , e , e .

Então, em ambos os casos nós aumentamos a quantidade entregue, mas deixamos fixas as informações já existentes. Por exemplo, na base de Aviação Comercial, nós não tínhamos acesso ao ano de entrega dos aviões, mas tínhamos acesso ao país de destino. Portanto, nós repetimos os países de destino para cada quantidade aumentada, e geramos aleatoriamente o ano de entrega dos aviões (prezando o período em que o modelo de aeronave foi fabricado). Exemplo: . Geramos:

|  |  |  |
| --- | --- | --- |
| Quantidade entregue | Ano de entrega | País de destino |
| 1 | 2006 | Austrália |
| 1 | 2007 | Austrália |
| 1 | 2008 | Austrália |

O mesmo procedimento foi realizado para a base de dados de Aviação Executiva, sendo os dados faltantes o país de destino e os dados preexistentes que foram repetidos foram os anos de entrega.

- Fluxos de Dados:

Diagrama

Descrição gerada automaticamente

3.1 Descrição do Fluxograma:

- Coleta de Dados:

- Inicia o processo de coleta de dados de diversas fontes na internet.

- Inserção em Planilha:

- Os dados coletados são inseridos em uma planilha para organização.

- Rotina em Python (ETL):

- Utiliza-se uma rotina em Python para realizar as operações de Extração, Transformação e Carga (ETL) nos dados armazenados na planilha. Operações ETL incluem:

- Extração: Os dados são lidos da planilha.

- Transformação: Aplicam-se transformações conforme necessário (limpeza, manipulação de dados).

- Carga: Os dados transformados são carregados nas tabelas do MySQL.

- Conexão MySQL com Power BI:

- Realiza-se a conexão entre o MySQL e o Power BI para possibilitar o desenvolvimento de relatórios.

- Desenvolvimento de Relatórios no Power BI:

- Utiliza-se o Power BI para criar relatórios interativos com base nos dados carregados no MySQL.

- Análise de Dados:

- Exploração e interpretação dos dados para obter insights valiosos.

3.2 Ferramentas Específicas Utilizadas durante o Processo:

A seguinte rotina em Python utiliza a biblioteca SQLAlchemy para realizar operações de ETL (Extração, Transformação e Carga) em dados armazenados em arquivos CSV para tabelas MySQL.

- Importação de Bibliotecas:

- `create\_engine` e `pd` (Pandas): Funções do SQLAlchemy e Pandas, respectivamente, para criação de engine de banco de dados e manipulação de dados tabulares.



- Configurações do Banco de Dados:

- `db\_config`: Dicionário contendo informações como host, usuário, senha e nome do banco de dados.

Texto

Descrição gerada automaticamente

- Conexão ao MySQL usando SQLAlchemy:

- `engine`: Cria uma instância de engine para gerenciar a conexão com o banco de dados.

- Mapeamento de Arquivos CSV para Tabelas:

- `mapeamento\_arquivos\_tabelas`: Dicionário que associa nomes de arquivos CSV a nomes correspondentes de tabelas MySQL.



- Loop sobre Arquivos e Tabelas:

- A rotina percorre cada arquivo e tabela correspondente, lê os dados do arquivo CSV usando Pandas e os insere no MySQL. Essa rotina realiza a ETL, garantindo a integridade e consistência dos dados no MySQL, além de fornecer informações detalhadas sobre eventuais erros encontrados durante o processo.

Texto

Descrição gerada automaticamente

4. Recomendações para Aprimorar a Confiabilidade e Qualidade dos Dados:

- Adotar Boas Práticas de Coleta:

- Implementar validações de formato e fazer verificações de consistência durante a entrada de dados.

- Considerar a adoção de técnicas de aprendizado de máquina para detecção automática de anomalias.

- Armazenar Dados:

- Normalizar o banco de dados para reduzir redundâncias e melhorar a consistência.

- Manter um plano de backup regular e testar procedimentos de recuperação.

- Limpar Dados:

- Utilizar expressões regulares ou ferramentas específicas para corrigir problemas de formatação.

- Automatizar processos de limpeza sempre que possível para eficiência a longo prazo.

- Monitorar Continuamente:

- Implementar alertas automáticos para notificar sobre possíveis problemas nos dados.

- Realizar auditorias periódicas para garantir a aderência contínua às boas práticas.