**6ª Etapa do Desenvolvimento de Projeto de Business Intelligence – BI**

**Andressa Carneiro Rocha, Elias Nogueira Salgado, Fernando Queiroz de Paula, Pedro Henrique dos Santos Machado, Thales de Mattos Oliveira,**

**Victor Hugo da Silva Aguiar, Vinícius Duarte Oliveira Lage**

**Orientadora: Viviane Cristina Dias**

PUC Minas

Curso de Tecnologia em Banco de Dados

**1º Identificar quais transformações deverão ser realizadas.**

Buscou-se compreender a estrutura dos dados fornecidos pelo rBLH, disponíveis em relatórios em formato PDF intitulados "rBLH em Dados". Esses relatórios estão divididos em 4 arquivos, que abrangem os anos de 2022, 2021, 2020 e um maior, com o período de 2000 a 2019. A organização desses dados envolve tabelas específicas para cada estado, detalhando informações de coleta ao longo dos meses, além de cinco tabelas que consolidam dados por regiões e uma tabela para o Brasil como um todo.

A decisão inicial foi concentrar a análise em um período de 10 anos, considerado adequado para uma análise abrangente. Observou-se que seria pertinente focar nas páginas dos PDFs contendo dados estaduais entre 2013 e 2022, totalizando 270 páginas. Essa escolha teve como pressuposto otimizar a carga de dados, visto que dados regionais e globais podem ser derivados por cálculos a partir dos dados específicos dos estados.

Ao explorar os PDFs, identificou-se que as tabelas estavam representadas como imagens. Optou-se, então, por utilizar a ferramenta de OCR da AWS Amazon Textract, capaz de analisar até 1000 páginas por mês de forma gratuita. Durante esse processo, notou-se uma peculiaridade no relatório de 2013: o estado do Acre não registrou coleta em abril, resultando em uma linha a menos na tabela. Para padronização, decidiu-se adicionar uma linha com valores zerados para abril, garantindo 13 linhas em todas as tabelas.

Ao avaliar os dados, identificou-se quatro tipos principais: Texto, Data, Números Inteiros e Números Decimais. A partir do modelo construído nas etapas anteriores, identificamos os dados disponíveis e delineamos as medidas necessárias para análises futuras.

**2º Definir a ferramenta ETL que será utilizada.**

Decidiu-se por utilizar o Power BI, alavancando o Power Query para conduzir as fases de extração, transformação e carga (ETL) dos dados. A decisão de adotar o Power BI foi motivada por sua capacidade robusta de visualização e integração com o Power Query, que se destaca pela flexibilidade na manipulação de dados.

Antecipou-se não apenas a necessidade de uma preparação eficaz dos dados, mas também a importância de criar visualizações impactantes nas etapas subsequentes do projeto. O Power BI oferece uma interface intuitiva e recursos avançados para atender a esses requisitos, consolidando assim uma abordagem coesa entre as operações de ETL e a apresentação visual de dados.

**3º Planejar e executar os Jobs necessários para extração, transformação e carga no data warehouse.**

As etapas de extração, carregamento e transformação (ELT) dos dados foram realizadas da seguinte forma:

Foram baixados os quatro relatórios disponíveis para os anos de 2022, 2021, 2020 e o último, que compreende dados de 2000 a 2019 no endereço < https://rblh.fiocruz.br/serie-documentos>.

Posteriormente, empregou-se o LibreOffice Draw para selecionar especificamente as páginas dos PDFs contendo dados estaduais dos anos de 2013 a 2022, totalizando 270 páginas.

Para a extração efetiva dos dados contidos nessas páginas, a ferramenta OCR da AWS Amazon Textract foi utilizada. A eficácia dessa ferramenta foi evidenciada pela entrega segura de um arquivo .csv para cada página analisada, resultando em 270 arquivos nomeados como table-x.csv, onde x representa o número da página. Como forma de validação, confrontou-se os valores nos arquivos .csv com as imagens e não foram identificadas discrepâncias.

Dentro desse processo, foi tratada uma particularidade no relatório de 2013 referente ao estado do Acre, que não registrou coleta em abril. Para manter a padronização das tabelas, a linha referente a abril com valores zerados foi inserida, assegurando assim 13 linhas em todas as tabelas.

O próximo passo envolveu o carregamento desses dados no Power BI. Utilizando a opção "Obter Dados" a partir da pasta contendo os arquivos .csv, utilizando a opção "Combinar e Transformar Dados".

A screenshot of a computer

Description automatically generated

Essa escolha iniciou o Power Query, gerando consultas, parâmetros e scripts para combinar os dados.

A screenshot of a computer

Description automatically generated

A partir dessa opção, as transformações realizadas em um arquivo de exemplo foram replicadas de forma padronizada e automática para os demais arquivos.

A screenshot of a computer

Description automatically generated

Durante esse processo, foram executadas transformações específicas utilizando linguagem M para adequar o arquivo de exemplo, garantindo que contivesse apenas os dados extraídos dos PDFs.

**linguagem M:**

let

Fonte = Csv.Document(Parâmetro1,44,"",ExtraValues.Ignore,65001),

#"Outras Colunas Removidas" = Table.SelectColumns(Fonte,{"Column1"}), // Remover colunas sem dados

#"Primeiras Linhas Mantidas" = Table.FirstN(#"Outras Colunas Removidas",13), // Manter apenas as colunas com dados relevantes

#"Dividir Coluna por Delimitador" = Table.SplitColumn(#"Primeiras Linhas Mantidas", "Column1", Splitter.SplitTextByDelimiter(",'", QuoteStyle.Csv), {"Column1.1", "Column1.2", "Column1.3", "Column1.4", "Column1.5", "Column1.6", "Column1.7", "Column1.8", "Column1.9", "Column1.10", "Column1.11", "Column1.12", "Column1.13", "Column1.14"}), // Dividir as colunas a partir do delimitador ,'

#"Valor Substituído" = Table.ReplaceValue(#"Dividir Coluna por Delimitador","'","",Replacer.ReplaceText,{"Column1.1"}), // Retirar os caracteres ' da primeira coluna

#"Valor Substituído1" = Table.ReplaceValue(#"Valor Substituído",",","",Replacer.ReplaceText,{"Column1.14"}), // Retirar os caracteres , da última coluna

#"Cabeçalhos Promovidos" = Table.PromoteHeaders(#"Valor Substituído1", [PromoteAllScalars=true]) // Promover a primeira linha para se tornar cabeçalho

in

#"Cabeçalhos Promovidos"

A screenshot of a computer

Description automatically generated

Adicionalmente, duas tabelas do Excel foram carregadas, uma contendo a dimensão localidade, contendo os estados e regiões, e outra a dimensão tempo, contendo os meses e anos estudados..

Na consulta final, as transformações iniciais foram aplicadas automaticamente:

**linguagem M:**

let

Fonte = Folder.Files("C:\PUC - TBD\Eixo 3\rBLH\Produção\2013-2022 csv"),

#"Arquivos Ocultos Filtrados1" = Table.SelectRows(Fonte, each [Attributes]?[Hidden]? <> true),

#"Invocar Função Personalizada1" = Table.AddColumn(#"Arquivos Ocultos Filtrados1", "Transformar Arquivo", each #"Transformar Arquivo"([Content])),

#"Colunas Renomeadas1" = Table.RenameColumns(#"Invocar Função Personalizada1", {"Name", "Nome da Origem"}),

#"Outras Colunas Removidas1" = Table.SelectColumns(#"Colunas Renomeadas1", {"Nome da Origem", "Transformar Arquivo"}),

#"Coluna de Tabela Expandida1" = Table.ExpandTableColumn(#"Outras Colunas Removidas1", "Transformar Arquivo", Table.ColumnNames(#"Transformar Arquivo"(#"Arquivo de Amostra"))),

A screenshot of a computer

Description automatically generated

As últimas transformações foram aplicadas na consulta final, incluindo a alteração de tipos de dados, extração de mês e ano, a mesclagem com as consultas Localidade e Tempo, e, por fim, a reordenação lógica das colunas.

**linguagem M:**

#"Coluna Duplicada1" = Table.DuplicateColumn(#"Coluna de Tabela Expandida1", "Nome da Origem", "Nome da Origem - Copiar"), //Coluna duplicada para gerar IdEstatisticasLeiteHumano

#"Colunas Renomeadas" = Table.RenameColumns(#"Coluna Duplicada1",{{"Nome da Origem", "Estado"}, {"Nome da Origem - Copiar", "IdEstatisticasLeiteHumano"}}),

#"Valor Substituído1" = Table.ReplaceValue(#"Colunas Renomeadas","table-","",Replacer.ReplaceText,{"Estado"}), //Texto retirado para sobrar apenas os números dos arquivos

#"Valor Substituído2" = Table.ReplaceValue(#"Valor Substituído1",".csv","",Replacer.ReplaceText,{"Estado", "IdEstatisticasLeiteHumano"}), //Texto retirado para sobrar apenas os números dos arquivos

#"Tipo Alterado" = Table.TransformColumnTypes(#"Valor Substituído2",{{"Atend. em Grupo", Int64.Type}, {"Atend. Individual", Int64.Type}, {"Visita Domiciliar", Int64.Type}, {"Doadoras", Int64.Type}, {"Receptores", Int64.Type}, {"Produção cadastradas no mês", Int64.Type}, {"Acidez Dornic", Int64.Type}, {"Crematócrito", Int64.Type}, {"Exame microbiológico", Int64.Type}, {"Leite Humano Recebido", type number}, {"Leite Humano Transferido", type number}, {"Leite Humano Distibuído", type number}, {"Leite Humano Coletado", type number}, {"Estado", Int64.Type}, {"Mês / Ano", type date}}), //Tipo de dados alterados

#"Coluna Condicional Adicionada1" = Table.AddColumn(#"Tipo Alterado", "IdLocalidade", each if [Estado] > 243 then [Estado] - 243 else if [Estado] > 216 then [Estado] - 216 else if [Estado] > 189 then [Estado] - 189 else if [Estado] > 162 then [Estado] - 162 else if [Estado] > 135 then [Estado] - 135 else if [Estado] > 108 then [Estado] - 108 else if [Estado] > 81 then [Estado] - 81 else if [Estado] > 54 then [Estado] - 54 else if [Estado] > 27 then [Estado] - 27 else [Estado] ), //Números dos arquivos reduzidos para menor de 27, visto que os ids de estados vão de 1 a 27.

#"Coluna Personalizada Adicionada" = Table.AddColumn(#"Coluna Condicional Adicionada1", "IdTempo", each Text.Combine({Date.ToText([#"Mês / Ano"], "yyyy"), Date.ToText([#"Mês / Ano"], "MM")}), type text), //idTempo adicionado

#"Colunas Removidas" = Table.RemoveColumns(#"Coluna Personalizada Adicionada",{"Estado", "Mês / Ano"}), // Reordenação para conferência

#"Consultas Mescladas" = Table.NestedJoin(#"Colunas Removidas", {"IdLocalidade"}, Localidade, {"IdLocalidade"}, "Localidade", JoinKind.LeftOuter), //Consulta mesclada para obter a dimensão localidade.

#"Consultas Mescladas1" = Table.NestedJoin(#"Consultas Mescladas", {"IdTempo"}, Tempo, {"IdTempo"}, "Tempo", JoinKind.LeftOuter),

// Coluna tempo removida para atender ao modelo.

#"Colunas Reordenadas" = Table.ReorderColumns(#"Consultas Mescladas1",{"IdEstatisticasLeiteHumano", "IdLocalidade", "IdTempo", "Atend. em Grupo", "Atend. Individual", "Visita Domiciliar", "Doadoras", "Receptores", "Leite Humano Coletado", "Leite Humano Distibuído", "Leite Humano Transferido", "Leite Humano Recebido", "Exame microbiológico", "Crematócrito", "Acidez Dornic", "Produção cadastradas no mês"}) //Consulta mesclada para obter a dimensão tempo.

in

#"Colunas Reordenadas"

A screenshot of a computer

Description automatically generated

Após essas etapas, as transformações foram aplicadas, e os dados foram carregados na tabela "Dados" no Power BI, além das tabelas Localidade e Tempo.

A screenshot of a computer

Description automatically generated

Com a estrutura de dados organizada, foram criadas medidas usando a ferramenta "Medida Rápida", calculando indicadores Entradas, Saídas, Produção Líquida Local e Total, além de métricas relacionadas à média de leite coletado, distribuído e disponível, e a razão entre doadoras e receptores.

A screenshot of a computer

Description automatically generated

Ao final desse processo, os dados foram preparados e estão prontos para gerar visualizações significativas no Power BI, conforme o modelo apresentado nas etapas anteriores.

A screenshot of a computer

Description automatically generated