Prof. Carlos Helano / Daniela Araújo

Objetivo

Desenvolver um projeto que contemple todo o processo de implantação de um BI. Mesmo não utilizando de todos os recursos e ferramentas necessárias para a implantação de um BI, este projeto visa proporcionar a vivência e o aprendizado de como seria a implantação de um projeto de BI. O escopo do projeto envolve a coleta e extração dos dados (ETL), a modelagem de um Data Warehouse (DW) com seus indicadores e a apresentação dos dados (OLAP).

Em geral, as ferramentas utilizadas no mercado para projetos de BI exigem um grande poder de processamento e armazenamento de dados. Além disso, muitas delas são ferramentas proprietárias que exigem licença de uso como: SAP, Oracle, Analysis Server / SQL Server (Microsoft). Algumas ferramentas podem ser utilizadas de forma gratuita como o Pentaho, mas este trabalho em si, não visa exatamente o aprendizado em cima do ferramental, pois exigiria uma infraestrutura e um tempo adequado para tal.

Descrição do Cenário

Para o desenvolvimento do projeto iremos investigar um conjunto de dados públicos disponibilizados pelo Ministério da Saúde. Estes dados dizem respeito internações hospitalares que acontecem em todo o território brasileiro pelo SUS (Sistema único de Saúde).

Toda internação hospitalar realizada pelo SUS recebe uma identificação chamada de AIH (Autorização de Internação Hospitalar). Estas internações são realizadas no SIHSUS (Sistema de Informações Hospitalares de SUS) pelos hospitais e estabelecimentos de saúde que atendem pelo SUS.

No link a seguir é possível saber um pouco da história do SIHSUS :
https://datasus.saude.gov.br/transferencia-de-arquivos/

Cada estabelecimento de saúde registrado no território brasileiro também recebe uma identificação conhecida como CNES (Cadastro Nacional de Estabelecimentos de Saúde). E cada registro de AIH feito por um destes estabelecimentos é enviado para o Ministério da Saúde para serem avaliados e assim aprovarem a verba para o pagamento dos custos da internação realizada por um determinado estabelecimento.

Cada registro de AIH (também chamada de conta de AIH), depois de aprovada ou rejeita pelo Ministério da Saúde é também disponibilizada ao público em geral como forma demonstrar a transparência no processo de pagamento destas contas para o cidadão. Nem toda informação é pública, pois existe também a preocupação de preservar a identidade do paciente. Por isso, não encontraremos por exemplo, o nome do paciente, nome do pai ou da mãe ou qualquer documento pessoal. Mas, encontraremos outras informações muito mais relevantes para um projeto de BI como: A quantidade de internações ocorridas em um determinado período, num determinado estado ou cidade, ou mesmo os montantes pagos para cada cidade ou estado. Da mesma forma é possível quantificar os procedimentos médicos que mais ocorreram e as doenças associadas (CID).

Organização

O trabalho será desenvolvido dupla e será dividido em 3 etapas.
Cada dupla será responsável por analisar pelo menos 1 unidade federativa (UF) no período de janeiro a dezembro nos anos de 2019, 2020, 2021, 2022, 2023 e 2024. Não utilizar: DF, SP, RJ e nem os estados do Norte.

Etapa 01 – Coleta dos Dados

A Coleta dos dados no site do DATASUS (Departamento de Informática do Sistema único de Saúde) deve ser feita através do site:
https://datasus.saude.gov.br/transferencia-de-arquivos/

Neste site é possível identificar os diversos sistemas mantidos pelo DATASUS. Como nosso escopo está voltado para as Internações (AIH) devemos escolher a opção SIHSUS. Nele poderemos fazer o download dos dados por Estado (UF) e no período desejado.

Outras bases de dados originadas de outros sistemas também se fazem necessário. Uma delas é a de estabelecimentos de saúde (CNES) também disponível no link acima e o download deve ser feita de forma semelhante ao SIHSUS.

Além destas, outras bases de dados se fazem necessário como a de Municípios (IBGE), Procedimentos de Saúde e CID (Código Internacional de Doenças), mas estas iremos disponibilizar em formato de texto (.CSV).

Os arquivos disponibilizados para o download estão num formato de compactação antigo com a extensão .DBC (Database Containe), e para termos acesso aos dados teremos que descompactar utilizando um software específico muito utilizado por especialistas, pesquisadores e estatísticos da área de saúde chamado TABWIN

TABWIN
http://www2.datasus.gov.br/DATASUS/index.php?area=060805&item=3

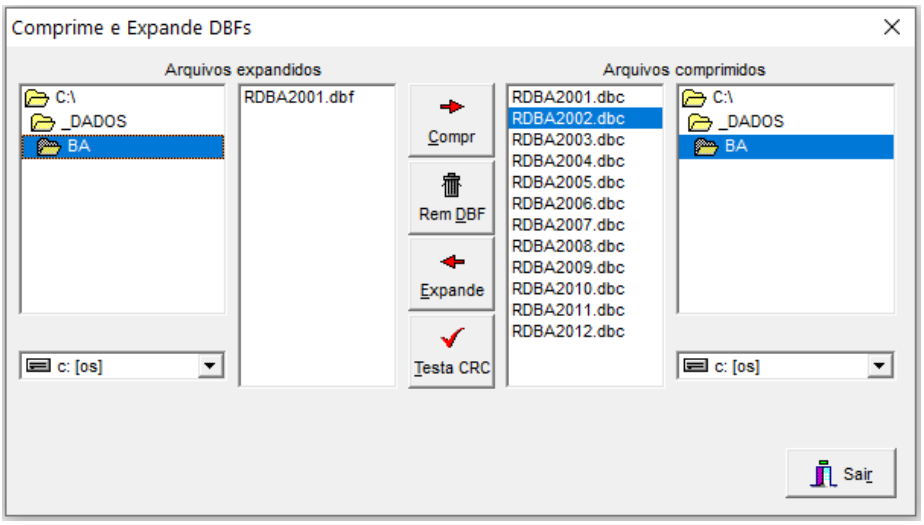
No link acima deveremos fazer o download do arquivo Tab415.zip.
Crie uma pasta C:\Tabwin e descompacte este arquivo zip.
Em seguida execute o aplicativo “TabWin415.exe”.

Dica: podem usar um gerenciador de download como o https://jdownloader.org/download/index

Para descompactar os arquivos de dados no formato DBC, faremos os seguintes procedimentos:

1 – Dentro do TABWIN no menu :

"Arquivo – Comprime/Expande.DBF"

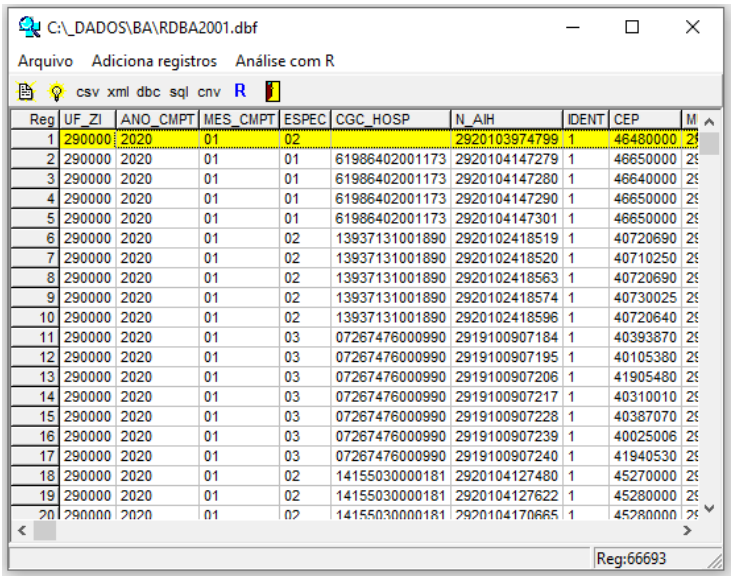


Nesta tela cada arquivo deve ser descompactado (botão Expande) para o formato DBF.

2 – Após a descompactação ainda não temos o formato desejado, pois os arquivos estão originalmente no formato .DBF, e assim ficamos limitados para tratar os dados em qualquer outro banco de dados. Assim sendo, devemos converter os arquivos .DBF para o formato de texto .CSV.

No próprio aplicativo TABWIN podemos fazer esta conversão no menu : Arquivo – Ver Arquivo. DBF

Em seguida, deve-se abrir o arquivo .DBF desejado e fazer a conversão para CSV (botão na barra de ferramentas). Após o clique no botão indicado, o arquivo CSV será criado na mesma pasta do arquivo de origem.



Reg	UF_ZI	ANO_CMPT	MES_CMPT	ESPEC	CGC_HOSP	N_AIH	IDENT	CEP	M
1	290000	2020	01	02		2920103974799	1	46480000	29
2	290000	2020	01	01	61986402001173	2920104147279	1	46650000	29
3	290000	2020	01	01	61986402001173	2920104147280	1	46640000	29
4	290000	2020	01	01	61986402001173	2920104147290	1	46650000	29
5	290000	2020	01	01	61986402001173	2920104147301	1	46650000	29
6	290000	2020	01	02	13937131001890	2920102418519	1	40720690	29
7	290000	2020	01	02	13937131001890	2920102418520	1	40710250	29
8	290000	2020	01	02	13937131001890	2920102418563	1	40720690	29
9	290000	2020	01	02	13937131001890	2920102418574	1	40730025	29
10	290000	2020	01	02	13937131001890	2920102418596	1	40720640	29
11	290000	2020	01	03	07267476000990	2919100907184	1	40393870	29
12	290000	2020	01	03	07267476000990	2919100907195	1	40105380	29
13	290000	2020	01	03	07267476000990	2919100907206	1	41905480	29
14	290000	2020	01	03	07267476000990	2919100907217	1	40310010	29
15	290000	2020	01	03	07267476000990	2919100907228	1	40387070	29
16	290000	2020	01	03	07267476000990	2919100907239	1	40025006	29
17	290000	2020	01	03	07267476000990	2919100907240	1	41940530	29
18	290000	2020	01	02	14155030000181	2920104127480	1	45270000	29
19	290000	2020	01	02	14155030000181	2920104127622	1	45280000	29
20	290000	2020	01	02	14155030000181	2920104170665	1	45280000	29

3 – Depois a conversão de todos os arquivos para .CSV é necessário gerar um banco de dados para que estes arquivos sejam importados. Este banco de dados representa os "dados brutos". Ou seja, não se trata de um banco de dados

relacional. Trata-se apenas da fonte original de dados para as próximas etapas do projeto.
O banco de dados a ser gerado é de livre escolha (Sql Server, SQL Express, MySql ou outros).

4 – Para efeito de registro e checagem das informações iniciais alguns levantamentos devem ser feitos:

4.a – Quantos registros de AIH existem para cada mês e em cada estado?
4.b – Quantos registros de Procedimentos de Saúde foram importados?
4.c – Quantos registros de CID foram importados?
4.d – Quantos registros de estabelecimentos (CNES) foram importados?

Outra forma de fazer todo este processo é utilizando Python com a biblioteca Pandas.

Etapa 02 – Modelagem do DW

Nesta etapa devemos criar uma nova base de dados com a modelagem dimensional para medir os indicadores que desejamos. Nesta etapa deve-se pensar como devemos estruturar as tabelas fato e dimensão.

Segue abaixo alguns indicadores interessantes para serem considerados:

1 – Quantidade de AIHs por: Estado, Município, CNES, Ano, Mês, Procedimento, CID.
2 – Valores pagos por: Estado, Município, CNES, Ano, Mês, Procedimento, CID
3 – Tempo de permanência de internação nas mesmas dimensões.

Na etapa anterior tratamos de parte do processo de ETL (Extract Transform Load) que compreendemos a extração/coleta dos dados. Agora teremos que trabalhar na transformação e carga dos dados para a base de dados onde modelamos o DW.

Devemos verificar com atenção os quantitativos levantados no item 4 da Etapa 01, pois agora teremos que alimentar novas tabelas com chaves e índices que estão relacionadas. Neste momento, a carga de dados pode gerar algumas inconsistências.

Etapa 03 – Apresentação

Construção de um painel (com 4 ou mais “telas”) com indicadores críticos, que dê uma visão rápida da sua performance, capaz de aceitar algum tipo de interatividade com o utilizador e que seja fácil de atualizar.

A indicação é de uso do Power BI, mas podem utilizar outras ferramentas, contanto que esteja disponível no momento da apresentação e não limite a entrega.

Deve ser elaborado um relatório do trabalho com os dados analisados.

A equipe deve apresentar o Dashboard e suas análises numa aula síncrona e postar no repositório do Teams da turma o documento com o relatório do trabalho.