

SEGUNDA ETAPA – ENRIQUECIMENTO

Os dados de churn de interesse principal da área de marketing, passarão pelo processo de enriquecimento de outras bases, além de tratamentos exploratórios visando eliminação de inconsistências através de tratamento de domínios, valores ausentes e outliers definidos pelo requisito do negócio.

O que deverá ser feito?

Exploração dos arquivos/tabelas: estado civil.json, escolaridade.txt que se encontram em C:\ProjetoChurn\Fonte e da tabela PERFIL_CLIENTE_VALORES cujos acessos são: user: etl e senha(sugerida): Pofd@123 no database: PROJETO_CHURN_AUDITORIA

Extração Ler dados de *churns* mensais do arquivo 1-CHURN_PERFIL.csv que se encontra em C:\ProjetoChurn\Destino

Transformação e Saídas requeridas (Load)

Saída1: Esta saída atenderá a área de marketing.

Pede-se gravar os dados no arquivo **2-CHURN_PERFIL_ENRIQUECIDO.csv** em C:\ProjetoChurn\Destino

O arquivo poderá ser substituído a cada nova execução no mês.

O Layout deverá ser:

- **Header:** com o nome das colunas abaixo listadas:

Name	Type	Format
id	Integer	#
genero	String	
idade	Integer	#
uf	String	
indcliente	Integer	#
anomes	Integer	#
estadocivil	String	
escolaridade	String	
mensalidade	Integer	#
participacao	Integer	#

- **Cabeçalho:** primeira linha nome das colunas
- **Separador das colunas:** ponto e vírgula (;)
- **Ecoding:** UTF-8

Enriquecimento

- Para os dados de estado civil e escolaridade devem-se trazer as descrições correspondentes aos respectivos códigos. Em caso de não existir deixar como null.
- Para o campo estado civil as descrições serão buscadas do arquivo “estado civil.json”, já para escolaridade no arquivo “escolaridade.txt”
- Os valores de mensalidades e coparticipação deverão ser buscados na tabela PERFIL_CLIENTE_VALORES, caso não haja, permitir deixar como null. Os valores deverão ser arredondados para inteiro.
- Não será permitido a perda de linhas nesta integração.

Transformações/limpeza sobre regras do negócio

- **UF** – não permite *null* na fonte e deverá somente ser aceitas UFs do nordeste - caso contrário substituir pela moda
- **Estado Civil** - Em caso de *null*, substituir pela moda
- **Idade** - Em caso de *null* e outliers (idades negativas e acima de 100 anos), substituir pela mediana.
- **Escolaridade** - Em caso de *null*, substituir pela moda
- **Mensalidade** - caso haja valores *null* - Fazer normalização do dado pela média condicional por idade. Serão considerados outliers valores acima de R\$4500,00, conforme faixa de valores vigentes.

De 0 a 19 - R\$ 250,00	De 20 a 25 – R\$ 550,00
De 26 a 30 – R\$ 675,00	De 31 a 40 – R\$ 750,00
De 41 a 50 – R\$ 875,00	De 51 a 60 – R\$ 1035,00
De 61 a 75 – R\$ 1500,00	Mais de 76 – R\$ 1850,00
- **Participação** - Em caso de *null* e outliers (considera outliers 4 * desvio padrão), substituir pela mediana

Anexo

Consulta usada step PERFIL_CLIENTE_TRIMESTRE

```
SELECT
  ID IDTBL
, ROUND(VLR_MENSAL,0) VLR_MENSAL
, ROUND(VLR_COPARTICIPACAO,0) VLR_COPARTICIPACAO
FROM PERFIL_CLIENTE_VALORES
ORDER BY ID
```

Help PDI

<https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/10.0.x/mk-95pdia003/pdi-transformation-steps>

UTF-8: é um tipo de codificação Unicode (padrão internacional para a representação e manipulação de texto em diferentes sistemas de escrita ao redor do mundo). UTF-8 é o formato de codificação mais amplamente utilizado e compatível com a maioria dos sistemas e linguagens. Ele é uma codificação variável que usa de 1 a 4 bytes para representar cada caractere.