PROJETO: TRATA CHURN

PRIMEIRA ETAPA - AUDITORIA

Os dados de churn serão de interesse principal ***da área de marketing****, porém o foco da primeira entrega será separar os dados que realmente representam clientes inativos (Churn) dos demais identificados com algum problema.*
Os dados com problemas serão enviados para as áreas de: diagnóstico de sistemas e auditoria interna, conforme especificação técnica a seguir.

O que deverá ser feito?

Exploração do arquivo SIS_CRM_TB_CHURN.txt

Extração Ler dados de *churns* mensais da operadora do arquivo SIS_CRM_TB_CHURN.txt

Transformação e Saídas requeridas (Load)

Saída1: Esta saída atenderá a área de auditoria interna e irá capturar clientes ativos (indcliente =1). Estes dados deverão ser gravados na tabela CLIENTE_CHURN_AUDITA, que está no database PROJETO_CHURN_AUDITORIA, cujos acessos são: user: etl e senha(sugerida): Pofd@123.

Os dados na tabela serão tratados pela área responsável.

A tabela deverá ser truncada a cada execução mensal.

Nome da coluna	Tipo	Comp...	.	..	Não nulo	Identidade
123 ID_CLI	bigint	8		.	[v]	[ ]
ABC CPF_CLI	varchar	11			[v]	[ ]
ABC NOME_COMPLET...	varchar	100			[v]	[ ]
123 IND_CLIENTE	int	4		.	[v]	[ ]
123 ANO_MES	int	4		.	[v]	[ ]
DATA_INSERCAO	datetime	8		.	[v]	[ ]
ABC GENERO	varchar	1			[v]	[ ]

Transformações/limpeza sobre regras do negócio

- CPFs repetidos não serão aceitos (somente permitido uma ocorrência de cada)
- Identidades gêneros, aqui tratado somente como gênero, deverão estar no padrão: Masculino(M), Feminino(F), Agênero(A) Bigênero(B) Outros (Não-binário Gênero-fluido etc.) (O) => **expressões regulares(REGEX) usadas no anexo deste documento**
- Todos os dados deverão estar em maiúsculo e com grafia correta
- Criar a coluna ANOMES no formato YYYYMM
- Criar a coluna Nome completo, sendo nome e sobrenome separados por espaço

- Data de inserção do dado no destino – data hora do Sistema Operacional
- Demais colunas do destino carga direta
- Demais processos de limpeza deverão ser avaliados, conforme requerido pelo schema destino

Saída2: Esta saída atenderá a área de diagnósticos do CRM.

Pede-se gravar os CPFs duplicados na planilha **CPFS_DUPLICADOS.xlsx** em C:\ProjetoChurn\Destino

A planilha poderá ser substituída a cada nova execução no mês.

O Layout deverá ser:

- Primeira linha com o nome e ordem das colunas abaixo listadas:

  ID mês ano cpf nome sobrenome

Transformações/limpeza sobre regras do negócio

- Manter dados brutos - carga direta sem tratamento

Saída3: Esta saída atenderá a área de Marketing.

Pede-se gravar os clientes inativados (indcliente<>1) no arquivo 1-CHURN_PERFIL.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
ID	Integer
genero	String
estado_civil	Integer
idade	Integer
uf	String
escolaridade	Integer
indcliente	Integer
anoMes	Integer

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

Transformações/limpeza sobre regras do negócio

- Indicador de cliente, se nulos, deverão ser zerados
- CPFs repetidos não serão aceitos (somente permitido uma ocorrência de cada)
- Identidades gêneros, aqui tratado somente como gênero, deverão estar no padrão: Masculino(M), Feminino(F), Agênero(A) Bigênero(B) Outros (Não-binário Gênero-fluido etc.) (O)
- Criar a coluna ANOMES no formato YYYYMM conforme colunas Ano Mês
- Todos os dados deverão estar em maiúsculos e com grafia correta
- Demais colunas do destino carga direta
- Demais processos de limpeza deverão ser avaliados, conforme requerido pelo schema destino

Anexo

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.

REGEX: Step PDI (Replace String – Trata_cpf_genereo)

Para cpf:

```
[-.]
```

Para gênero:

```
\b(masculino|masc|m)\b

\b(feminino|fem|f)\b

\b(big[êe]nero|bi)\b

\b(ag[êe]nero|a)\b

\b(outros)\b
```