

Bootcamp IGTI: Desenvolvedor Business Intelligence**Trabalho Prático**

Módulo 03	Aplicações em ETL
------------------	--------------------------

Objetivos

Exercitar os seguintes conceitos vistos em sala de aula:

- ✓ Modelagem dimensional.
- ✓ ETL.
- ✓ Ferramenta ETL.
- ✓ Processo de carga.

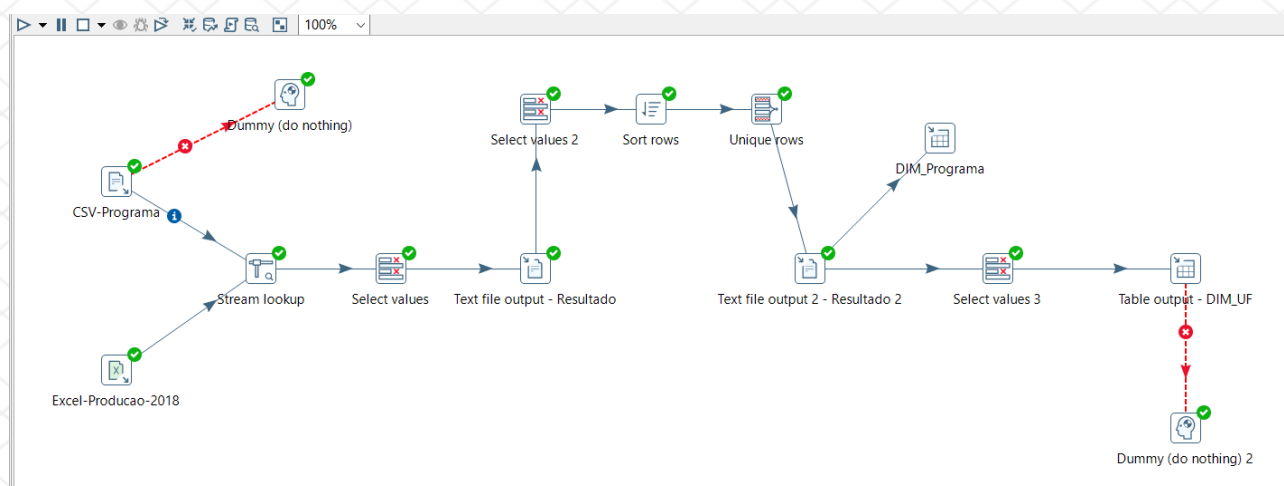
Enunciado

Vamos utilizar dois arquivos de entrada com informações da Capes, nos quais são feitas avaliações da pós-graduação stricto sensu. Essa avaliação tem como objetivo a certificação da qualidade da pós-graduação brasileira, bem como a identificação de assimetrias regionais e de áreas estratégicas do conhecimento. Ela é orientada pela Diretoria de Avaliação/CAPES, e realizada com a participação da comunidade acadêmico-científica por meio de consultores ad hoc. Tem como pilares a formação pós-graduada de docentes para todos os níveis de ensino e de profissionais de recursos humanos qualificados para o mercado, bem como o fortalecimento das bases científicas, tecnológicas e de inovação.

Estes dois arquivos servirão como entradas de dados para executarmos algumas atividades de transformação no Pentaho DTI.

Objetivos

Construiremos uma transformação que, ao final, alimentará uma tabela Dimensão chamada **dim-programa**.



Atividades

Processo de carga:

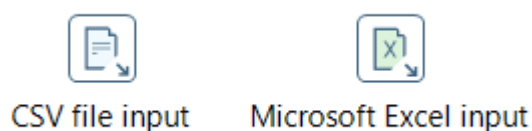
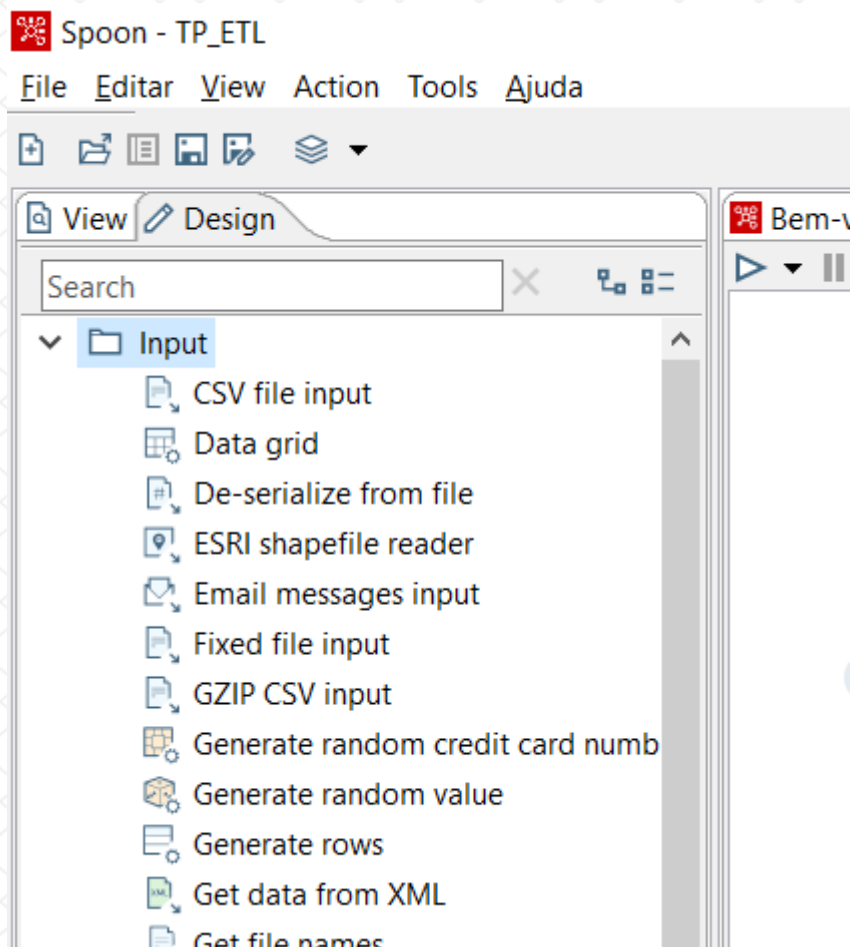
O primeiro passo é baixar os arquivos no link abaixo ou pegá-los na plataforma do IGTI:

<https://drive.google.com/drive/folders/1aWX2XKdbnCtH2UwnA3YsUcdZfwHe1Yr>

- *Producao-2018-Bibliografica-Artigo.xls.*
- *Programa-2018-Capes.csv.*

Para iniciar, abra o Spoon e crie uma nova transformação (*menu File | Novo | Transformação*).

Abra a categoria input e selecione os steps *Excel Input* e *CSV file Input*.



Edite o step Excel input com os seguintes parâmetros:

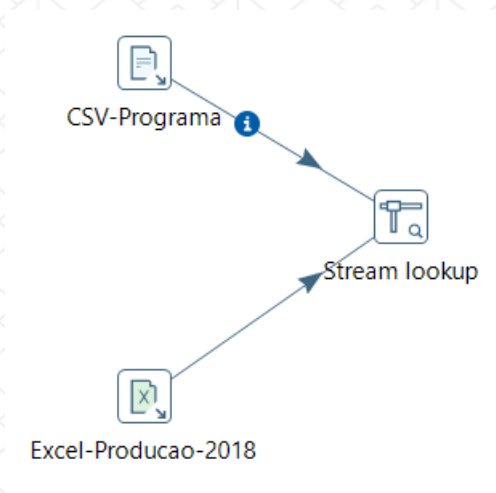
- Aba Files
 - File or directory: informe o arquivo *Producao-2018-Bibliografica-Artigo.xls*. Para ter certeza de que o arquivo foi localizado, clique no botão *show filename* para se certificar de que ele está sendo exibido.
- Aba Sheets
 - Clique no botão *Get sheetname* e escolha a planilha desejada.

- Aba Content
 - Certifique-se que o campo *Header* esteja marcado.
 - Se o arquivo fosse gravado em Linux (não é nosso caso), você precisaria mudar o campo *Encoding* para UTF-8.
- Aba Fields
 - Clique no botão *Get fields from header now* e veja todos os campos disponíveis.
 - Dê uma olhada nos dados que serão extraídos do arquivo clicando no botão *Preview rows*. Clique *Ok* e salve a transformação.

Edite o step CSV file input com os seguintes parâmetros:

- Filename: localize o arquivo *Programa-2018-Capes.csv* com o botão *Navegar*.
- Delimiter: informe: “;” (ponto e vírgula).
- Desmarque a opção *Lazy conversion*.
- Clique no botão *Obtem campos* e veja os campos que serão lidos.
- Clique no botão *Preview* para visualizar uma amostra dos dados.
 - Retire o símbolo da moeda (R\$) da propriedade *Currency*. Basta apagar essa informação em todos os atributos.
- Confirme se o *Header row present* está marcado. Isso indica que o arquivo tem um header com o nome dos campos.

Adicione o step *Stream lookup* e faça as ligações conforme figura abaixo.



Selecione o step *Stream lookup* com os seguintes parâmetros:

- Lookup step: escolha o step do CSV.
- Clique nos botões *Get fields* e *Get lookup fields*. O lookup fará a ligação dos arquivos pela chave comum. É uma ligação de 1 para N, ou seja, um programa tem N publicações, e a ligação entre os arquivos se dá pelo campo **CD_PROGRAMA_IES**. Temos esse campo nos dois arquivos: é o identificador do Programa.

Stream lookup

Step name
Stream lookup

Lookup step
CSV-Programa

The key(s) to look up the value(s):

#	Field	LookupField
1	CD_PROGRAMA_IES	CD_PROGRAMA_IES

Specify the fields to retrieve :

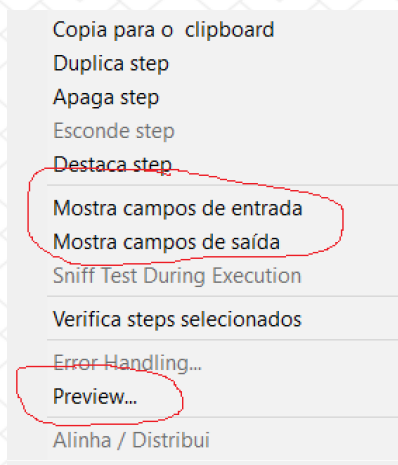
#	Field	New name	Default	Type
1	AN_BASE			Integer
2	NM_GRADE_AREA_CONHECIMENTO			String
3	NM_AREA_CONHECIMENTO			String
4	NM_AREA_BASICA			String
5	NM_SUBAREA_CONHECIMENTO			String
6	NM_ESPECIALIDADE			String
7	CD_AREA_AVALIACAO			Integer
8	NM_AREA_AVALIACAO			String
9	CD_ENTIDADE_CAPES			Integer
1..	CD_ENTIDADE_EMEC			String
1..	SG_ENTIDADE_ENSINO			String
1..	NM_ENTIDADE_ENSINO			String
1..	NM_REGIAO			String
1..	SG_UF_PROGRAMA			String
1..	NM_MUNICIPIO_PROGRAMA_IES			String
1..	NM_MODALIDADE_PROGRAMA			String

☒ Preserve memory (costs CPU)
☐ Key and value are exactly one integer field
☐ Use sorted list (i.s.o. hashtable)

Help
OK
Cancela
Get Fields
Get lookup fields

- Na grade de cima, remova todos os campos, deixando apenas o campo CD_PROGRAMA_IES.
- Na grade de baixo, deixe os campos que deseja que sejam mostrados a partir da junção.

Clique com o botão direito em cima do step Stream lookup e escolha as opções de visualização indicadas abaixo:



Suponha que não precisamos de todos os campos vindos da junção. Para isso, é necessário inserir o step *Select values*, fazendo a ligação necessária.

- Aba *Select & After* mostra os campos após você clicar no botão *Get fields to select*.
- Aba *Remove* mostra os campos após você clicar no botão *Get fields to remove*. Os campos que permanecerem nessa lista serão removidos.

Insira o step Text file output com os seguintes parâmetros:

- Aba *File*
 - *Filename*: caminho onde irá salvar o arquivo + nome do arquivo txt.
- Aba *Fields*
 - Clique no botão *Obtém campos* e veja os campos que serão gravados.

Você pode fazer alterações diretamente na grade. Clique no botão *Minimal width* e veja que o step fornece um formato padrão para os campos.

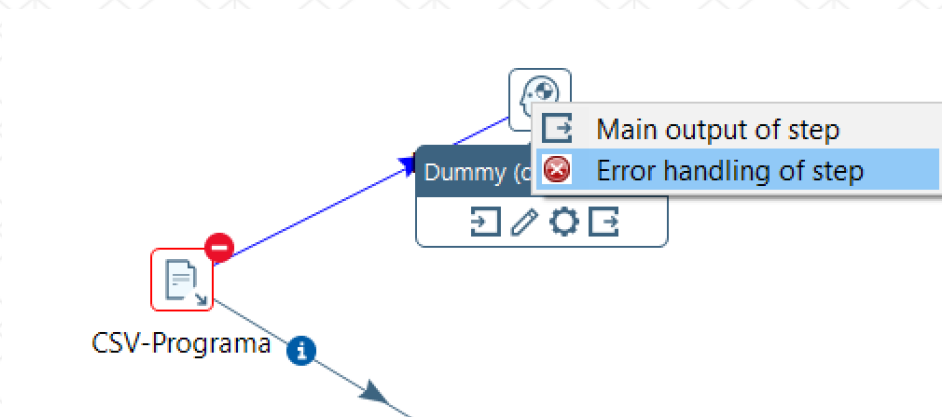
Rode o arquivo de transformação e veja se arquivo texto foi gravado no diretório indicado.

Observe que, inicialmente, ocorrerá um erro referente à linha 878. Na realidade, teremos erro em duas linhas. Isso porque há uma tentativa de converter valor no campo CD_CONCEITO_PROGRAMA de String para Integer.

```
2020/07/26 15:32:25 - CSV-Programa.0 - ERROR (version 9.0.0.0-423, build 9.0.0.0-423 from 2020-01-31 04:53.04 by buildguy) : Erro inesperado
2020/07/26 15:32:25 - CSV-Programa.0 - ERROR (version 9.0.0.0-423, build 9.0.0.0-423 from 2020-01-31 04:53.04 by buildguy) : org.pentaho.di.core.excep
2020/07/26 15:32:25 - CSV-Programa.0 - 
2020/07/26 15:32:25 - CSV-Programa.0 - There were 1 conversion errors on line 878
2020/07/26 15:32:25 - CSV-Programa.0 - 
2020/07/26 15:32:25 - CSV-Programa.0 - Unexpected conversion error while converting value [CD_CONCEITO_PROGRAMA String] to an Integer
2020/07/26 15:32:25 - CSV-Programa.0 - 
2020/07/26 15:32:25 - CSV-Programa.0 - CD_CONCEITO_PROGRAMA String : couldn't convert String to Integer
2020/07/26 15:32:25 - CSV-Programa.0 -
```

Você deve tratar o erro na transformação para que o processo não seja interrompido.

Insira o step *Dummy (do nothing)* e ligue o step CSV ao step *Dummy*. Na ligação selecione *Error handling for step*. Aparecerá uma caixa onde você deve aceitar a distribuição pelo botão *Distribuir*.



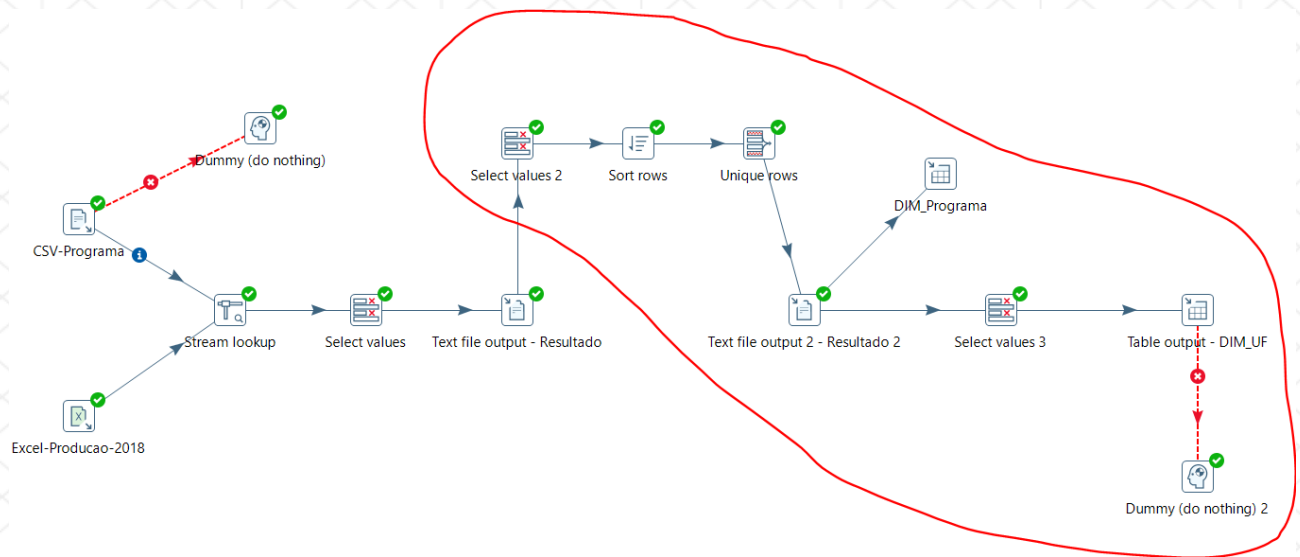
Rode novamente e veja que será concluído com sucesso.

Clique com o botão direito no step *Dummy* e selecione *Preview*. Veja as linhas que deram erro.

Rows of step: Dummy (do nothing) (2 rows)

#	AN_BASE	NM_GRADE_AREA_CONHECIMENTO	NM_AREA_CONHECIMENTO	NM_AREA_BASICA	NM_SUBAREA_CC
1	2018	MULTIDISCIPLINAR	CIÊNCIAS AMBIENTAIS	CIÊNCIAS AMBIENTAIS	NÃO SE APLICA
2	2018	CIÊNCIAS HUMANAS	CIÊNCIA POLÍTICA	CIÊNCIA POLÍTICA	NÃO SE APLICA

Agora é com você. Veja a figura abaixo e termine o trabalho prático.



Observe o conteúdo de cada step.

Select values 2

A saída desse step é:

Step name: Select values 2												
Fields:												
#	Fieldname	Type	Length	Precision	Step origin	Storage	Mask	Currency	Decimal	Group	Trim	Comments
1	CODIGO-IES	String	-	-	Select values	normal					nenhum	
2	NM_PROGRAMA_IES	String	-	-	Excel-Producao-2018	normal					nenhum	
3	SG_ENTIDADE_ENSINO	String	-	-	Excel-Producao-2018	normal					nenhum	
4	NM_ENTIDADE_ENSINO	String	-	-	Excel-Producao-2018	normal					nenhum	
5	NM_REGIAO	String	12	-	Stream lookup	normal		R\$,	.	nenhum	
6	SG_UF_PROGRAMA	String	2	-	Stream lookup	normal		R\$,	.	nenhum	
7	NM_MUNICIPIO_PROGRAMA_IES	String	21	-	Stream lookup	normal		R\$,	.	nenhum	
8	NM_MODALIDADE_PROGRAMA	String	12	-	Stream lookup	normal		R\$,	.	nenhum	

Sort rows

Sort rows

Nome do Step

Sort directory Navega...

TMP-file prefix

Sort size (rows in memory)

Free memory threshold (in %)

Compress TMP Files? ☐

Only pass unique rows? (verifies keys only) ☐

Fields :

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength
1	CODIGO-IES	S	N	N	0

Help

OK

Cancela

Obtem campos

Unique rows



Unique rows

Essa transformação tem como função remover linhas duplicadas do fluxo de entrada e filtrar apenas as linhas exclusivas para seguimento no fluxo.

Pré-Requisito: para entregar um resultado correto, o fluxo de entrada deve ser classificado em uma etapa anterior. Caso contrário, apenas as linhas duplas consecutivas serão analisadas e filtradas. Podemos utilizar a step *Sort rows* para isso.

linhas únicas

Nome do Step
Unique rows

Settings

Add counter to output? ☐ Counter field

Redirect duplicate row ☐ Error description

Fields to compare on (no entries means: compare complete row)

#	Fieldname	Ignore case
1	CODIGO-IES	N
2	NM_PROGRAMA_IES	N
3	SG_ENTIDADE_ENSINO	N
4	NM_ENTIDADE_ENSINO	N
5	NM_REGIAO	N
6	SG_UF_PROGRAMA	N
7	NM_MUNICIPIO_PROGRAMA_IES	N
8	NM_MODALIDADE_PROGRAMA	N

Help
OK
Cancela
Get

Text file output 2 – Resultado 2

Text file output

Nome do Step
Text file output 2 - Resultado 2

File
Content
Fields

Filename
D:\OneDrive\Aula\IGTI\ETL Bootcamp\TP e Desafio\TP\Resultado2

Pass output to servlet ☐

Create Parent folder ☒

Do not create file at start ☐

Accept file name from field? ☐

File name field

Extensão
txt

Include stepnr in filename? ☐

Include partition nr in filename? ☐

Include date in filename? ☐

Include time in filename? ☐

Specify Date time format ☐

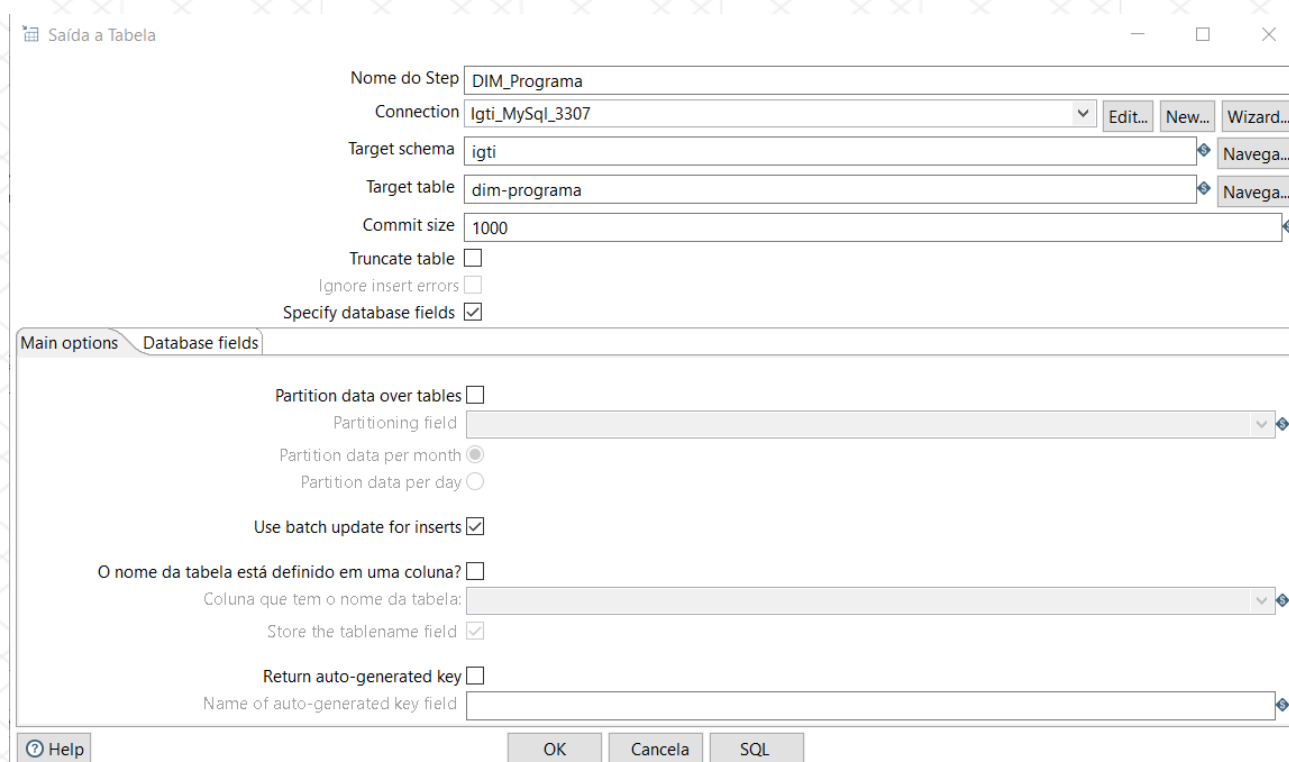
Date time format

Show filename(s)...

Add filenames to result ☒

Help
OK
Cancela

DIM_Programa (step Table output)



Saída a Tabela

Nome do Step: DIM_Programa

Connection: lgti_MySql_3307 [Edit... New... Wizard...]

Target schema: igti [Navega...]

Target table: dim-programa [Navega...]

Commit size: 1000

Truncate table: ☐

Ignore insert errors: ☐

Specify database fields: ☒

Main options Database fields

Partition data over tables: ☐

Partitioning field: []

Partition data per month: ☒

Partition data per day: ☐

Use batch update for inserts: ☒

O nome da tabela está definido em uma coluna?: ☐

Coluna que tem o nome da tabela: []

Store the tablename field: ☒

Return auto-generated key: ☐

Name of auto-generated key field: []

[?] Help [OK] [Cancela] [SQL]

O step *Table output* carrega dados em uma tabela do banco de dados. Esse step é equivalente ao operador SQL INSERT, e é uma solução quando você só precisa inserir registros. Se você quiser apenas atualizar linhas, use Update. Para executar os comandos *INSERT* e *UPDATE*, consulte a etapa Insert / Update.

Esta etapa fornece opções de configuração para uma tabela de destino e opções relacionadas ao desempenho, como *Commit Size* e *Use batch update* para inserções. Existem configurações de desempenho específicas para um tipo de banco de dados, que podem ser definidas nas propriedades JDBC da conexão com o banco de dados.

Use um gerenciador de Banco de Dados. No nosso caso é o *MySql*, mas você pode usar outro banco de dados se conseguir estabelecer a conexão.

Connection: conexão com o banco de dados.

Database Connection

General
Advanced
Options
Pooling
Clustering

Connection name:
Igti_MySql_3307

Connection type:
MySQL
Native Mondrian
Neoview
Netezza
Oracle
Oracle RDB
Palo MOLAP Server
Pentaho Data Services
PostgreSQL
Redshift
Remedy Action Request System
SAP ERP System

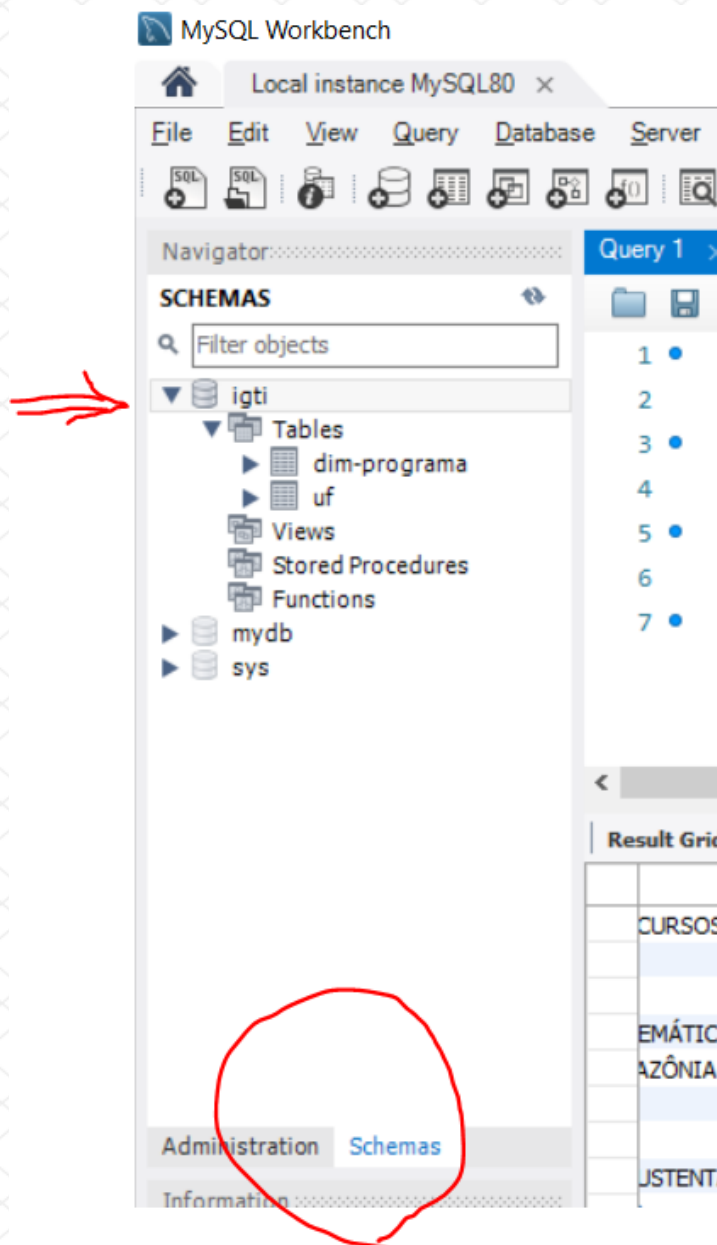
Access:
Native (JDBC)
ODBC
JNDI

Settings
Host Name:
localhost
Database Name:
igti
Port Number:
3307
Username:
root
Password:
•••••
☒ Use Result Streaming Cursor

Test
Feature List
Explore

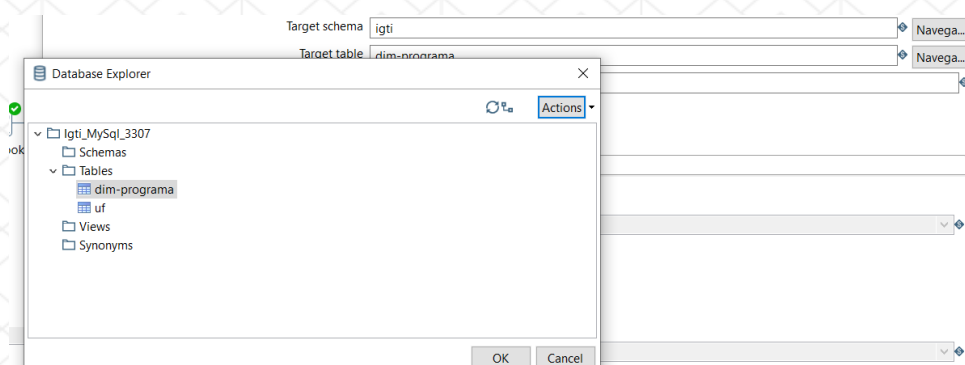
OK
Cancel

Target schema: nome do schema.

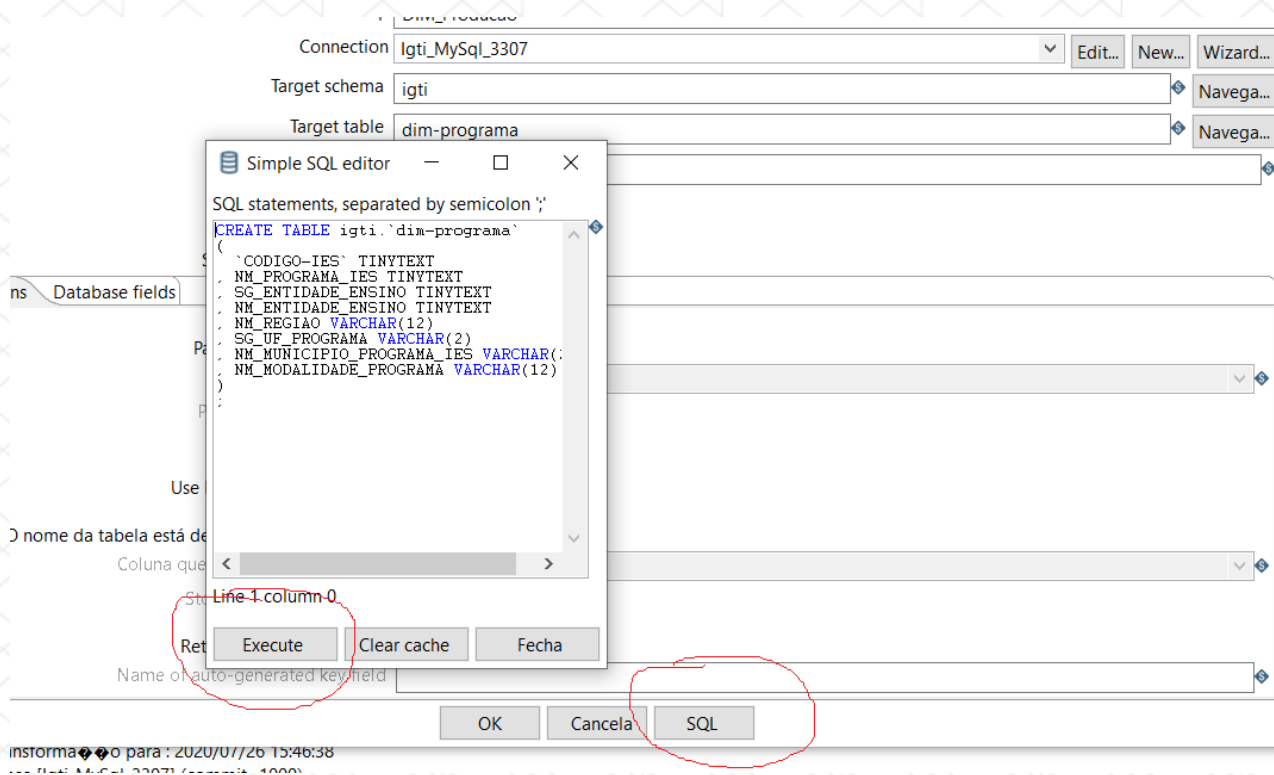


Target table: nome da tabela.

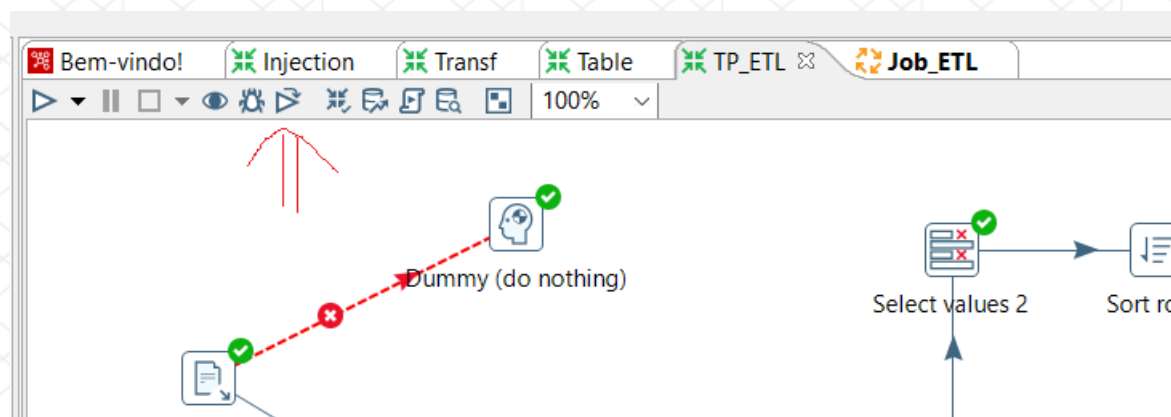
Para esse caso, informe o nome da tabela e clique no botão *Navegar*. Se a conexão estiver ok, ele vai mostrar o caminho no banco de dados.



Se a tabela não existir ainda, clique no botão *SQL*. Uma nova janela com o script de criação da tabela abrirá. Clique no botão *Execute*, e a tabela será criada.



Rode a transformação.



Clique no step *DIM_Producao* com o botão direito e selecione *Preview* para ver os dados.

Examine preview data

Rows of step: DIM_Programa (1000 rows)

#	CODIGO-IES	NM_PROGRAMA_IES	SG_ENTIDADE_ENSINO	NM_ENTIDADE_ENSINO	NM_REGIAO	SG_UF_PROGRAMA
1	10001018002P1	BIOLOGIA EXPERIMENTAL	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
2	10001018005P0	GEOGRAFIA	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
3	10001018009P6	PSICOLOGIA	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
4	10001018011P0	EDUCAÇÃO	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
5	10001018016P2	EDUCAÇÃO ESCOLAR	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
6	10001018039P2	DIREITOS HUMANOS E DESENVOLVIMENTO DA JUSTIÇA	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO
7	11001011001P8	ECOLOGIA E MANEJO DE RECURSOS NATURAIS	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC
8	11001011004P7	PRODUÇÃO VEGETAL	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC
9	11001011005P3	SAÚDE COLETIVA	FIOCRUZ	FUNDACAO OSWALDO CRUZ (FIOCRUZ)	NORTE	AC
1.	11001011006P0	CIÊNCIA, INOVAÇÃO E TECNOLOGIA PARA A AMAZÔNIA	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC
1.	11001011008P2	SANIDADE E PRODUÇÃO ANIMAL SUSTENTÁVEL NA AMAZÔNIA OCIDENTAL	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC
1.	11001011071P6	CIÊNCIA FLORESTAL	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC
1.	12001015002P7	QUÍMICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015006P2	FÍSICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015008P5	GEOCIÊNCIAS	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015012P2	INFORMÁTICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015014P5	SOCIEDADE E CULTURA NA AMAZÔNIA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015016P8	CIÊNCIAS FLORESTAIS E AMBIENTAIS	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
1.	12001015021P1	ENGENHARIA ELÉTRICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015023P4	HISTÓRIA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015025P7	CIÊNCIAS PESQUEIRAS NOS TRÓPICOS	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015027P0	SERVIÇO SOCIAL	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015032P3	CIÊNCIAS DA COMUNICAÇÃO	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015034P6	IMUNOLOGIA BÁSICA E APLICADA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM
2.	12001015036P9	PSICOLOGIA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM

Close Stop Get more rows

Rode um comando sql no Banco de Dados para verificar os dados que foram inseridos na tabela.

```
select * from igti.`dim-programa`;
```

MySQL Workbench

Local instance MySQL80 x

File Edit View Query Database Server Tools Scripting Help

Navigator

SCHEMAS

Filter objects

igti

Tables

dim-programa

uf

Views

Stored Procedures

Functions

mydb

sys

Administration Schemas

Information

Schema: igti

Query 1 x SQL File 5"

Limit to 1000 rows

```
1 select * from igti.`dim-programa`;
```

Result Grid

Filter Rows:

Export:

Wrap Cell Contents

Fetch rows:

CODIGO-IES	NM_PROGRAMA_IES	SG_ENTIDADE_ENSINO	NM_ENTIDADE_ENSINO	NM_REGIAO	SG_UF_PROGRAMA	NM_MUNICIPIO_PRC
10001018002P1	BIOLOGIA EXPERIMENTAL	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
10001018005P0	GEOGRAFIA	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
10001018009P6	PSICOLOGIA	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
10001018011P0	EDUCAÇÃO	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
10001018016P2	EDUCAÇÃO ESCOLAR	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
10001018039P2	DIREITOS HUMANOS E DESENVOLVIMENTO DA ...	UNIR	UNIVERSIDADE FEDERAL DE RONDÔNIA	NORTE	RO	PORTO VELHO
11001011001P8	ECOLOGIA E MANEJO DE RECURSOS NATURAIS	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC	RIO BRANCO
11001011004P7	PRODUÇÃO VEGETAL	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC	RIO BRANCO
11001011005P3	SAÚDE COLETIVA	FIOCRUZ	FUNDACAO OSWALDO CRUZ (FIOCRUZ)	NORTE	AC	RIO BRANCO
11001011006P0	CIÊNCIA, INOVAÇÃO E TECNOLOGIA PARA A A...	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC	RIO BRANCO
11001011008P2	SANIDADE E PRODUÇÃO ANIMAL SUSTENTÁVE...	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC	RIO BRANCO
11001011071P6	CIÊNCIA FLORESTAL	UFAC	UNIVERSIDADE FEDERAL DO ACRE	NORTE	AC	RIO BRANCO
12001015002P7	QUÍMICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM	MANAUS
12001015006P2	FÍSICA	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM	MANAUS
12001015008P5	GEOCIÊNCIAS	UFAM	UNIVERSIDADE FEDERAL DO AMAZONAS	NORTE	AM	MANAUS

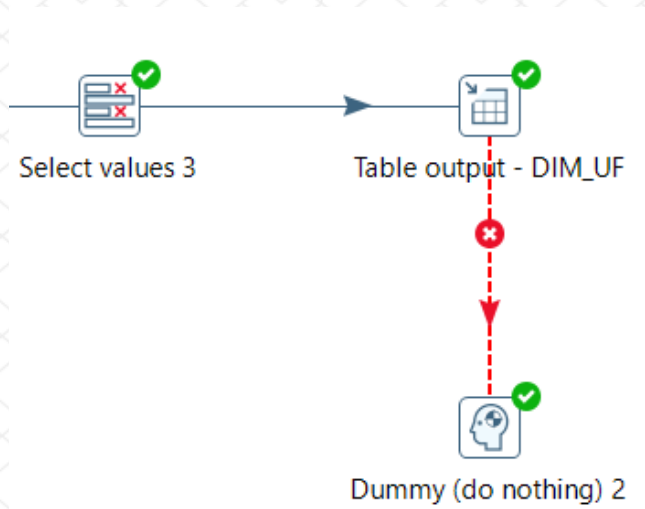
dim-programa 21 x

Essa parte é opcional. Você também tem liberdade para melhorar qualquer um dos steps.

Esse caso servirá para carregar uma dimensão UF a partir dos dados do arquivo de Resultado 2.

- Select values 3

- Table output (UF)
- Dummy (do nothing) 2



- Criar um arquivo de Job e linkar a transformação.

