

Bootcamp: Desenvolvedor Business Intelligence

Desafio do módulo

Módulo 3	Aplicações em ETL
-----------------	--------------------------

Atividades

Os alunos deverão desempenhar as seguintes atividades:

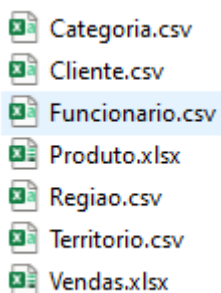
1. Utilizar o banco de dados MySQL e o Pentaho. É possível utilizar um gerenciador de banco de dados relacional de sua preferência.
2. Executar todo o processo de ETL no Pentaho conforme orientações.

Objetivos

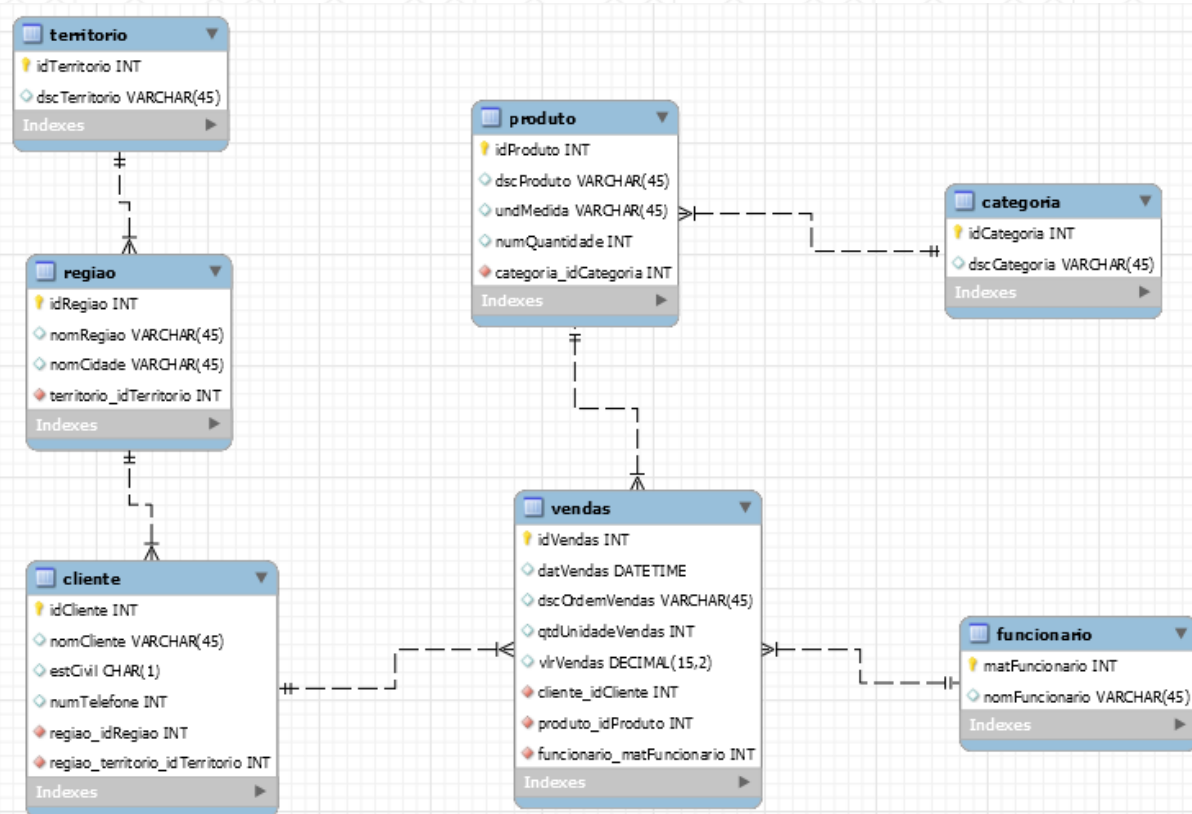
O objetivo desse exercício é fazer o processo completo do ETL no Pentaho, a partir de tabelas excel e arquivos csv.

Enunciado

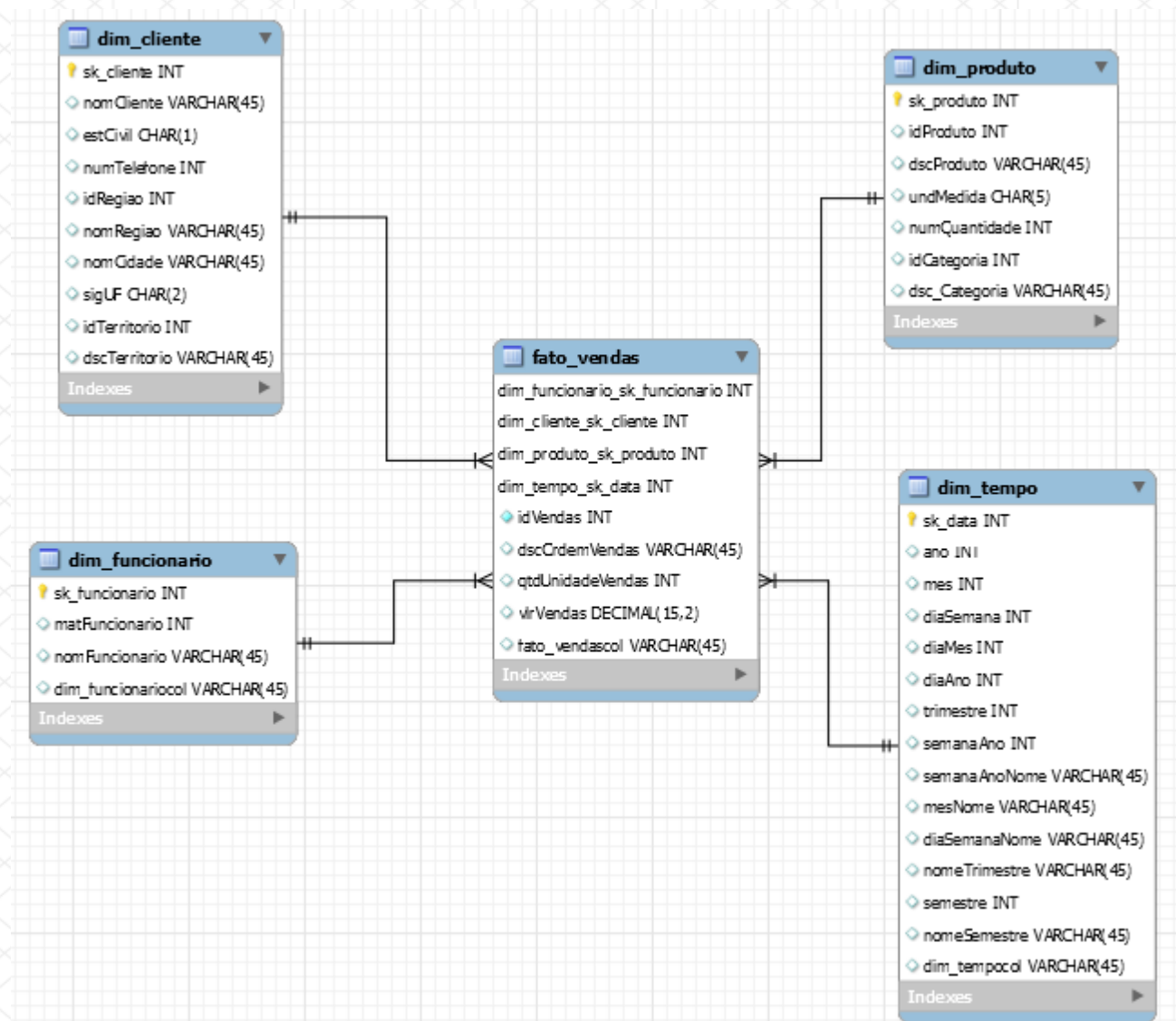
A partir de arquivos de dados do anexo “Origem” (csv e excel) que nos servirão como dados da origem, vamos modelar um DW.



Esses dados seguem a seguinte modelagem:

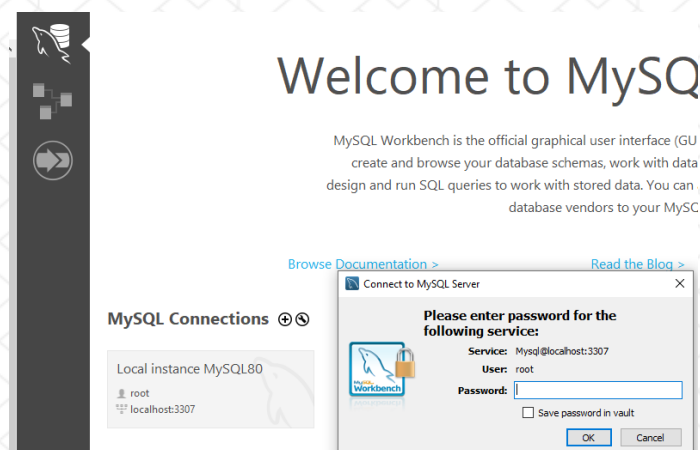


A partir desses dados de origem, o objetivo é modelar um DW como esquema abaixo:

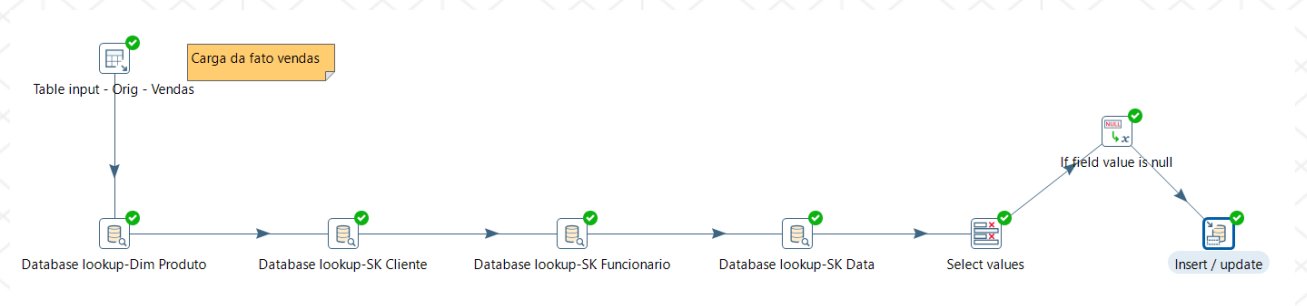
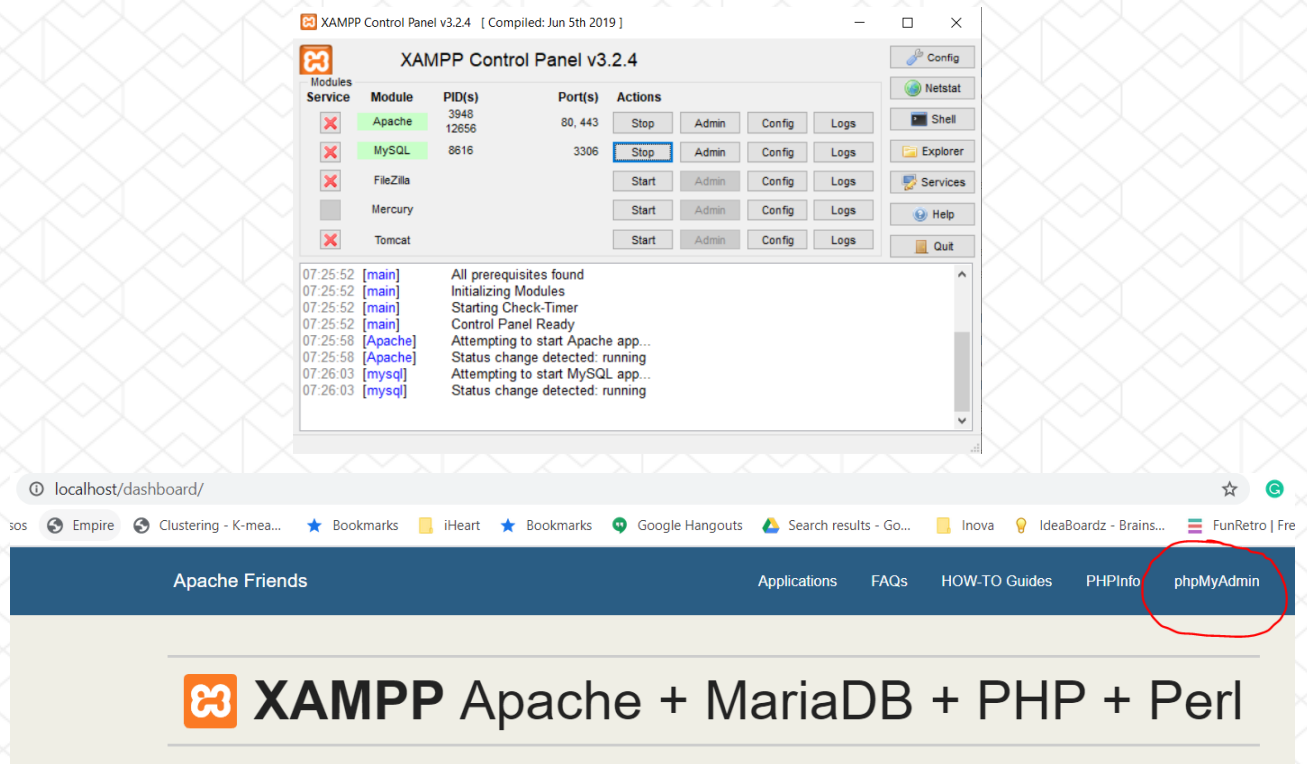


MySQL:

Estou utilizando o MySQL via Workbench, como o repositório para a origem.



O MySql via XAMPP/PHPAdmin servirá como o repositório para o DW.



Se preferir, pode utilizar somente uma instalação do MySql.

Precisamos de dois schemas diferentes, um para a Stage e outro para o DW. Criar os schemas no(s) MySql manualmente ou via comando sql:

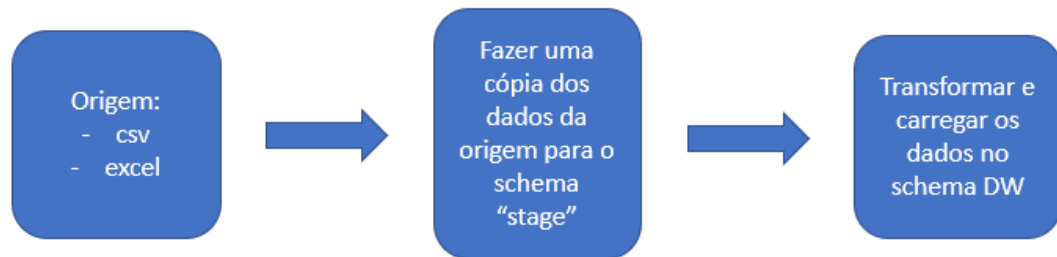
Exemplo:

No Workbench

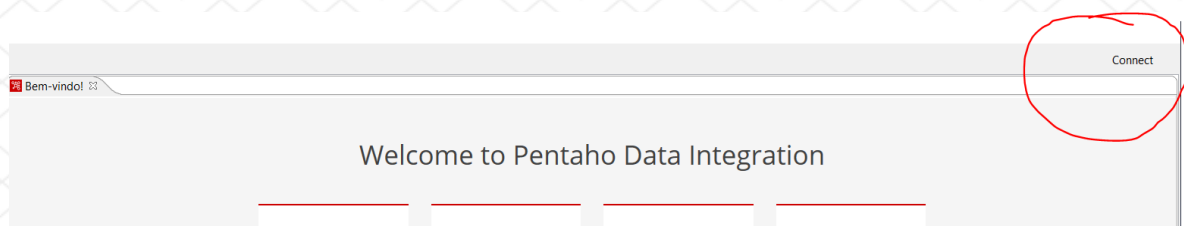
```
CREATE SCHEMA IF NOT EXISTS `stage` DEFAULT CHARACTER SET utf8 ;
```

No XAMPP Apache

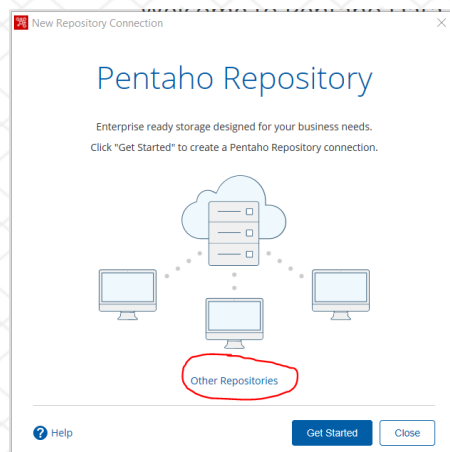
```
CREATE SCHEMA IF NOT EXISTS `dw` DEFAULT CHARACTER SET utf8 ;
```



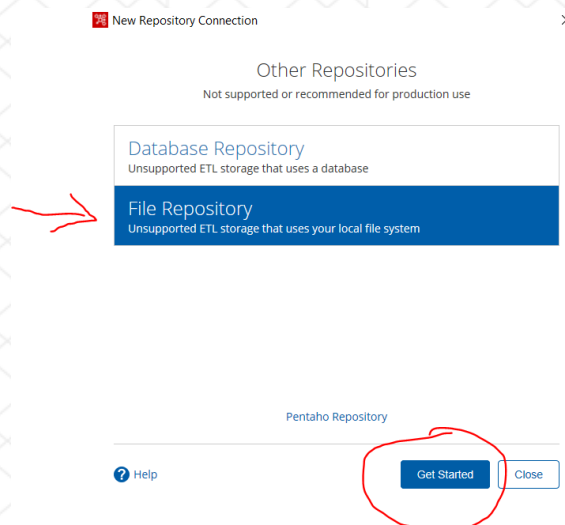
Vamos começar criando um repositório para o projeto do desafio.
Clique no botão conect.



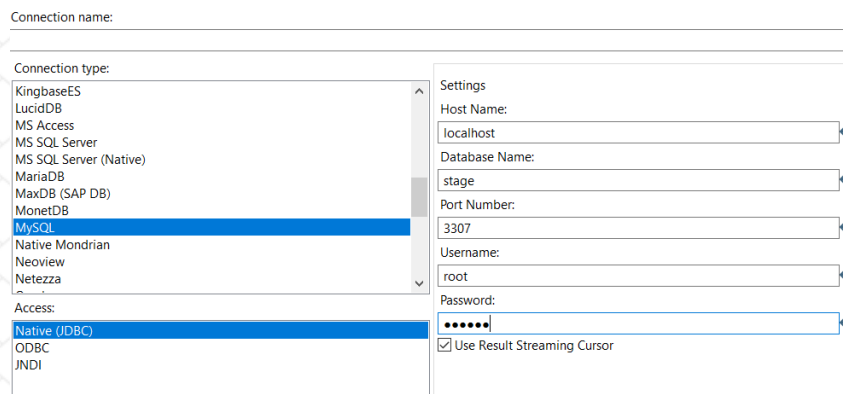
Clique em "Other Repositories":



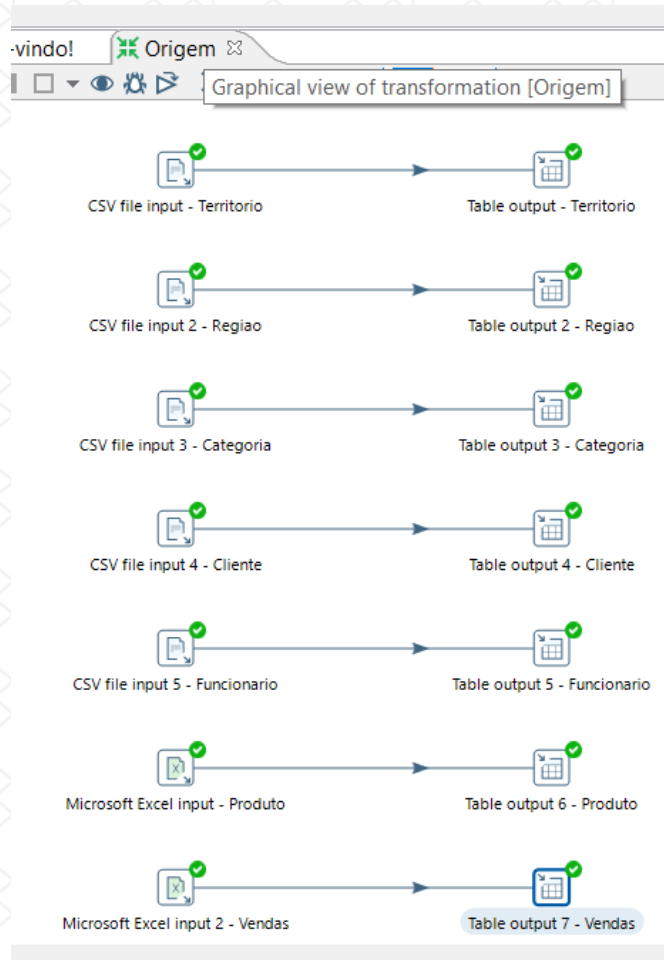
Selecione “File Repository” e clique em “Get Started”. Configure com o diretório de sua preferência para mandar o projeto salvo.



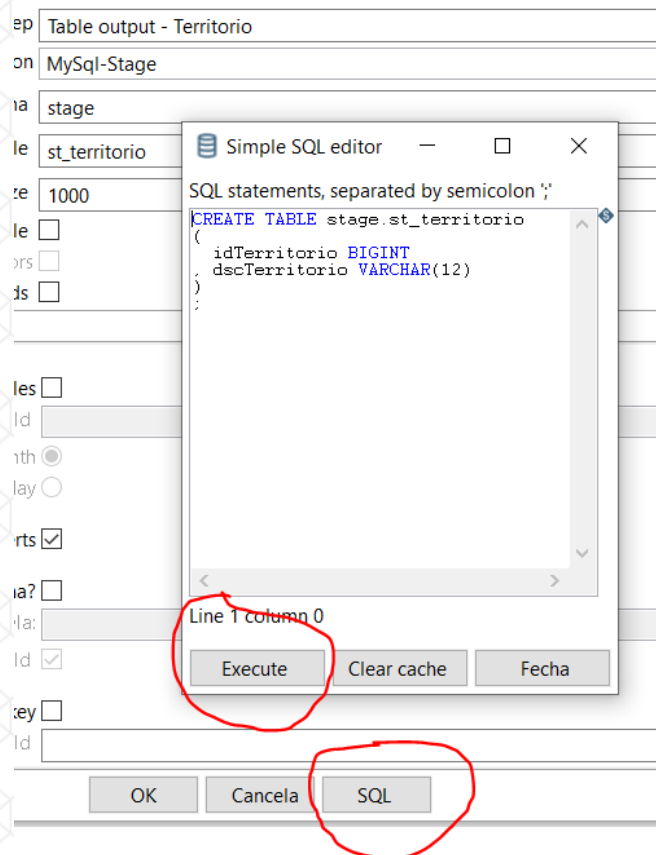
Para iniciar os trabalhos, vamos criar as conexões com os bancos MySQL.



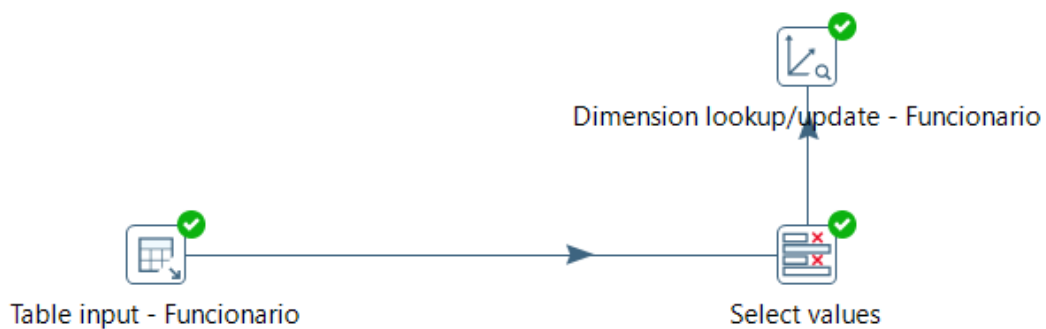
Vamos iniciar a construção das transformações, trazendo os dados da origem para a stage:



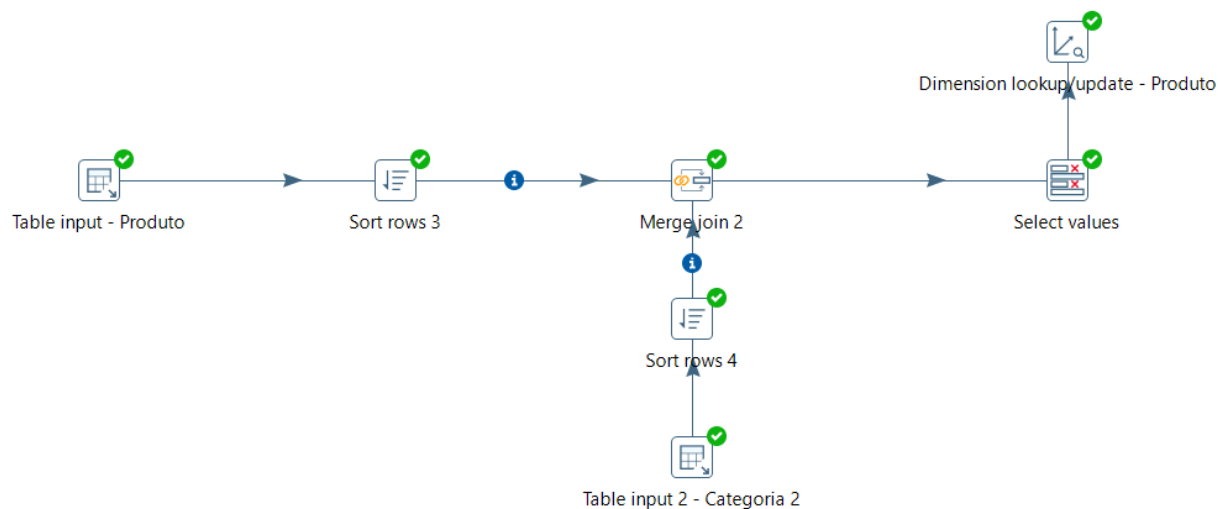
Uma dica, não é preciso criar as tabelas da stage e do DW nos bancos MySQL de forma manual. Você consegue fazer isso, por exemplo, pelo componente table output.



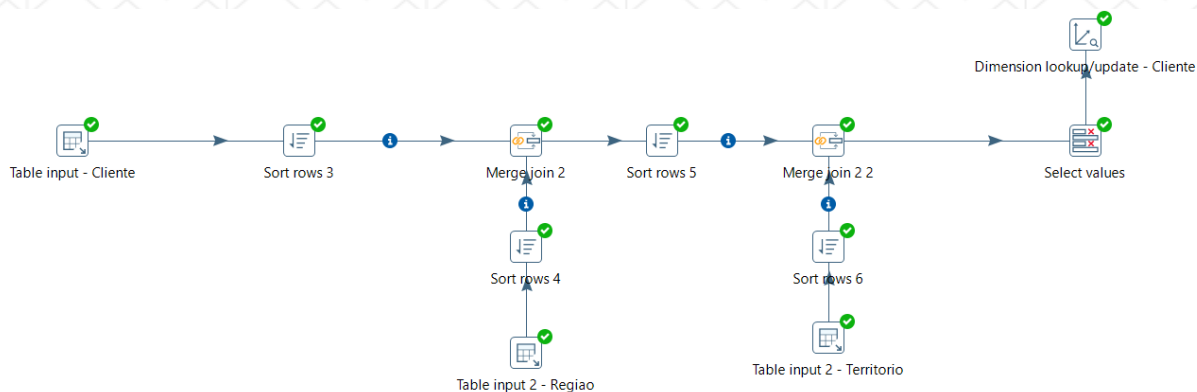
Carga para dimensão funcionário:



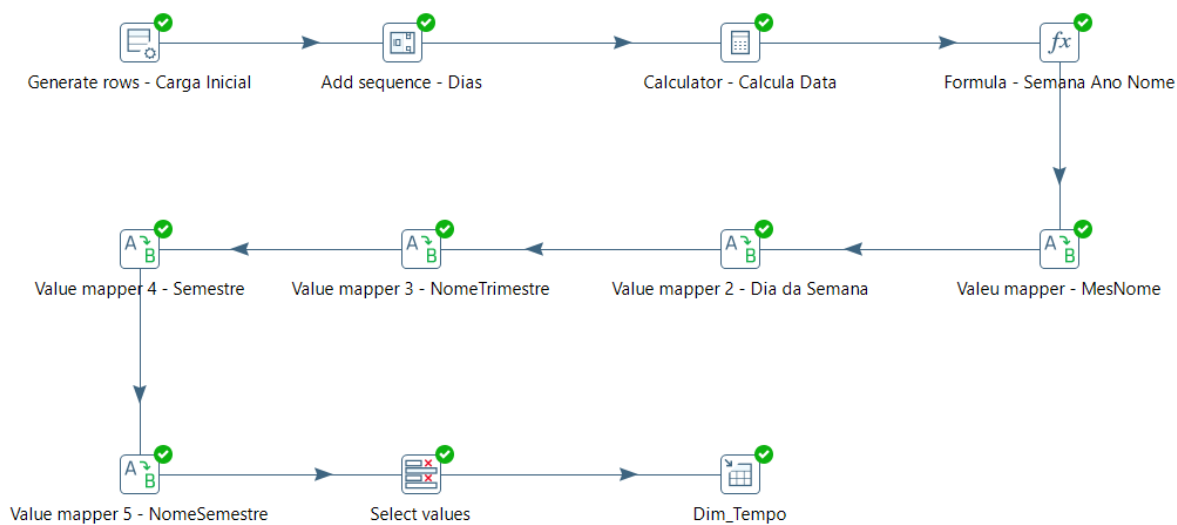
Carga para Dimensão Produto:



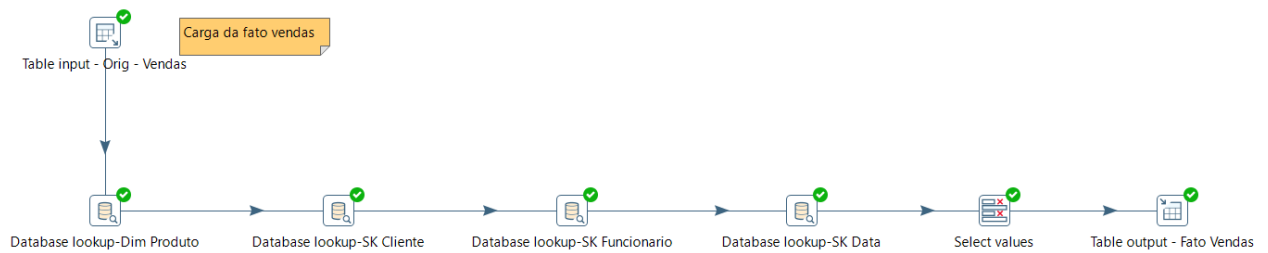
Carga para Dimensão Cliente:



Carga para Dimensão Tempo:



Carga da tabela Fato Venda:

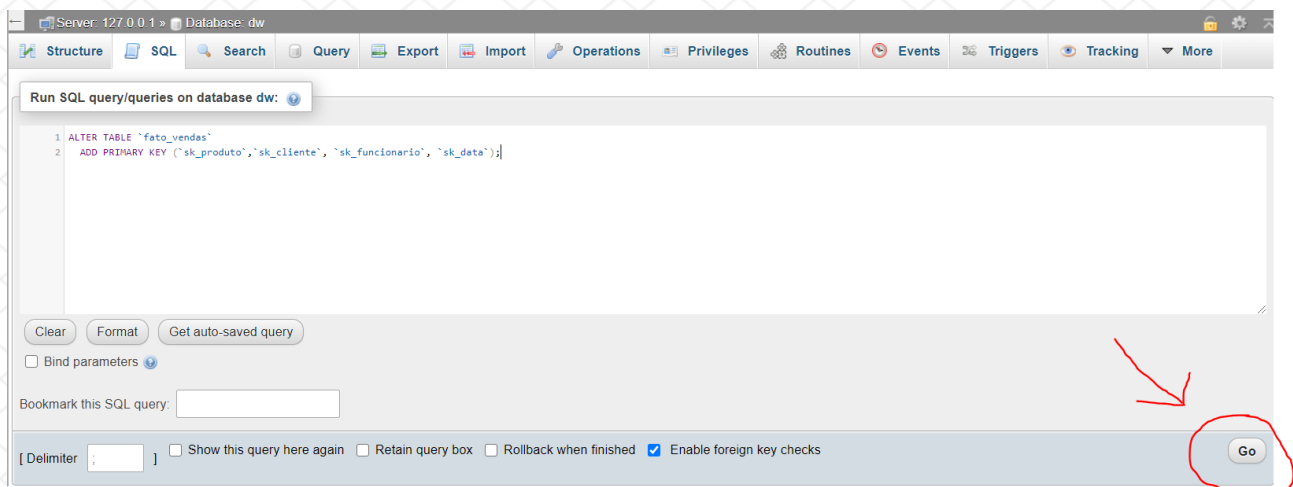


Repare pelo “preview” que nas últimas linhas da tabela fato a sk_produto está nula. Isso não deu erro porque a tabela fato não está com as surrogate keys das dimensões setada como chave. Vamos fazer um ajuste e ver o que acontece.

Rode o comando sql no banco de dados do schema onde está o DW.

```
ALTER TABLE `fato_vendas`
```

```
ADD PRIMARY KEY (`sk_produto`, `sk_cliente`, `sk_funcionario`, `sk_data`);
```



```
ALTER TABLE `fato_vendas` ADD PRIMARY KEY (`sk_produto`,`sk_cliente`,`sk_funcionario`,`sk_data`)
```

Warning: #1265 Data truncated for column 'sk_funcionario' at row 161

Warning: #1265 Data truncated for column 'sk_funcionario' at row 162

Warning: #1265 Data truncated for column 'sk_funcionario' at row 163

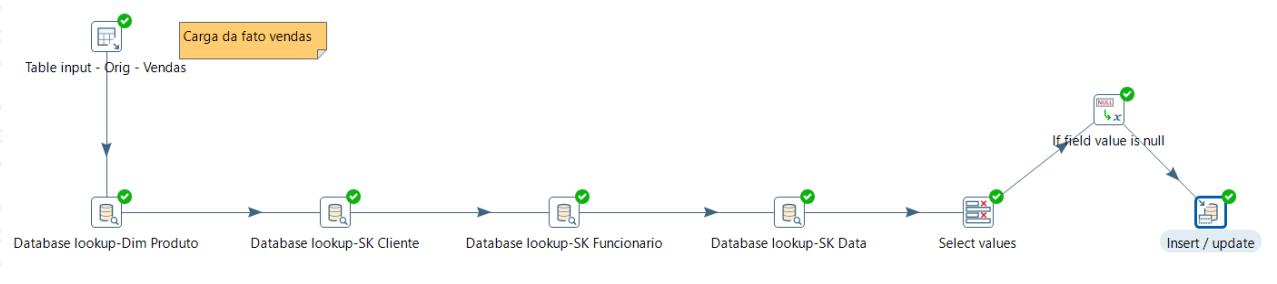
Warning: #1265 Data truncated for column 'sk_funcionario' at row 164

Warning: #1265 Data truncated for column 'sk_funcionario' at row 165

Agora que ligamos as chaves primárias na fato, tente rodar a carga da fato novamente. Você vai obter o erro: “Column ‘sk_funcionario’ cannot be null”.

```
2020/08/08 10:26:47 - Database lookup-SK Data.0 - Finished processing (I=100, O=0, R=100, W=100, U=0, E=0)
2020/08/08 10:26:47 - Select values.0 - Finished processing (I=0, O=0, R=165, W=165, U=0, E=0)
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - ERROR (version 9.0.0.0-423, build 9.0.0.0-423 from 2020-01-31 04:53:04 by buildguy) : Unexpected batch
2020/08/08 10:26:47 - Table output - Fato Vendas.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.co
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - Error updating batch
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - Column 'sk_funcionario' cannot be null
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - 
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - at org.pentaho.di.core.database.Database.createKettleDatabaseBatchException(Database.java:1434)
2020/08/08 10:26:47 - Table output - Fato Vendas.0 - at org.pentaho.di.core.database.Database.executeAndCommit(Database.java:1433)
```

Vamos resolver isso trocando o step “table output” por um step “insert update”.



Vamos configurar quem são as chaves na tabela e isso vai permitir fazer um update em caso de insert sem sucesso.

Reparem que precisaremos criar um registro nas dimensões com a Surrogate Key = -1. Vamos reservar o atributo com a SK = -1 para o caso de erro na carga.

Rode o comando sql no banco de dados do schema onde está o DW.

```
delete from `fato_vendas`;
```

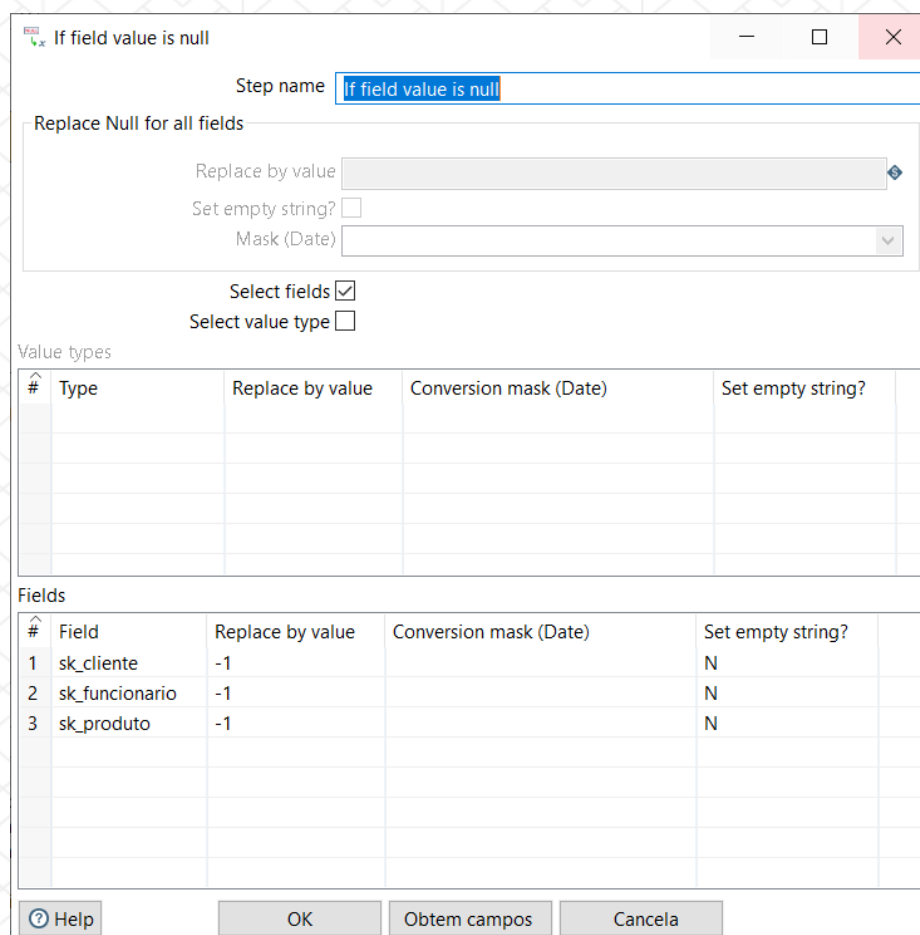
```
insert into `dim_produto` (`sk_produto`, `dscProduto`) values (-1, 'Erro');
```

```
insert into `dim_cliente` (`sk_cliente`, `nomCliente`) values (-1, 'Erro');
```

```
insert into `dim_funcionario` (`sk_funcionario`, `nomFuncionario`) values (-1, 'Erro');
```

Teremos que inserir um step “if field is null”. Configure conforme figura abaixo. No caso de termos valores nulos para as surrogate keys, vamos atribuir um valor -1 para elas.

Notaram que o valor -1 = erro está cadastrado nas dimensões? Isso vai nos ajudar a controlar os problemas nas chaves da tabela Fato.



Step name: If field value is null

Replace Null for all fields

Replace by value:

Set empty string? ☐

Mask (Date):

Select fields ☒

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)	Set empty string?

Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	sk_cliente	-1		N
2	sk_funcionario	-1		N
3	sk_produto	-1		N

Buttons: Help, OK, Obtem campos, Cancela

Configure o step “insert / update”, conforme figura abaixo.

Rode a carga e depois confira o resultado no MySQL. Verique que as últimas linhas tem a sk_funcionario com valores -1, ou seja, tem dados na tabela fato referenciando dados nas dimensões, porém esse dados na dimensão está ausente.

						idVenc	dscOrdemVendas	qtdUnidadeVenc	vlrVendas	sk_produ	sk_cliei	sk_funciona	sk_data
<input type="checkbox"/>	Edit	Copy	Delete	84	Ordem 84			300 3000.0		10	5	5	2020-01-05 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	126	Ordem 126			300 3000.0		24	64	17	2020-01-10 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	135	Ordem 135			1200 12000.0		25	67	23	2020-01-10 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	144	Ordem 144			1500 15000.0		26	34	23	2020-01-10 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	13	Ordem 13			400 4000.0		27	4	14	2020-01-02 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	89	Ordem 89			800 8000.0		27	63	10	2020-01-05 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	97	Ordem 97			1600 16000.0		28	41	2	2020-01-05 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	21	Ordem 21			1200 12000.0		28	51	6	2020-01-02 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	165	Ordem 165			100 8000.0		28	59	-1	2020-01-10 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	161	Ordem 161			600 6000.0		28	83	-1	2020-01-10 00:00:00
<input type="checkbox"/>	Edit	Copy	Delete	162	Ordem 162			700 7000.0		28	84	-1	2020-01-10 00:00:00

Chegamos ao final do nosso desafio, que é fazer o processo de ETL no Pentaho.