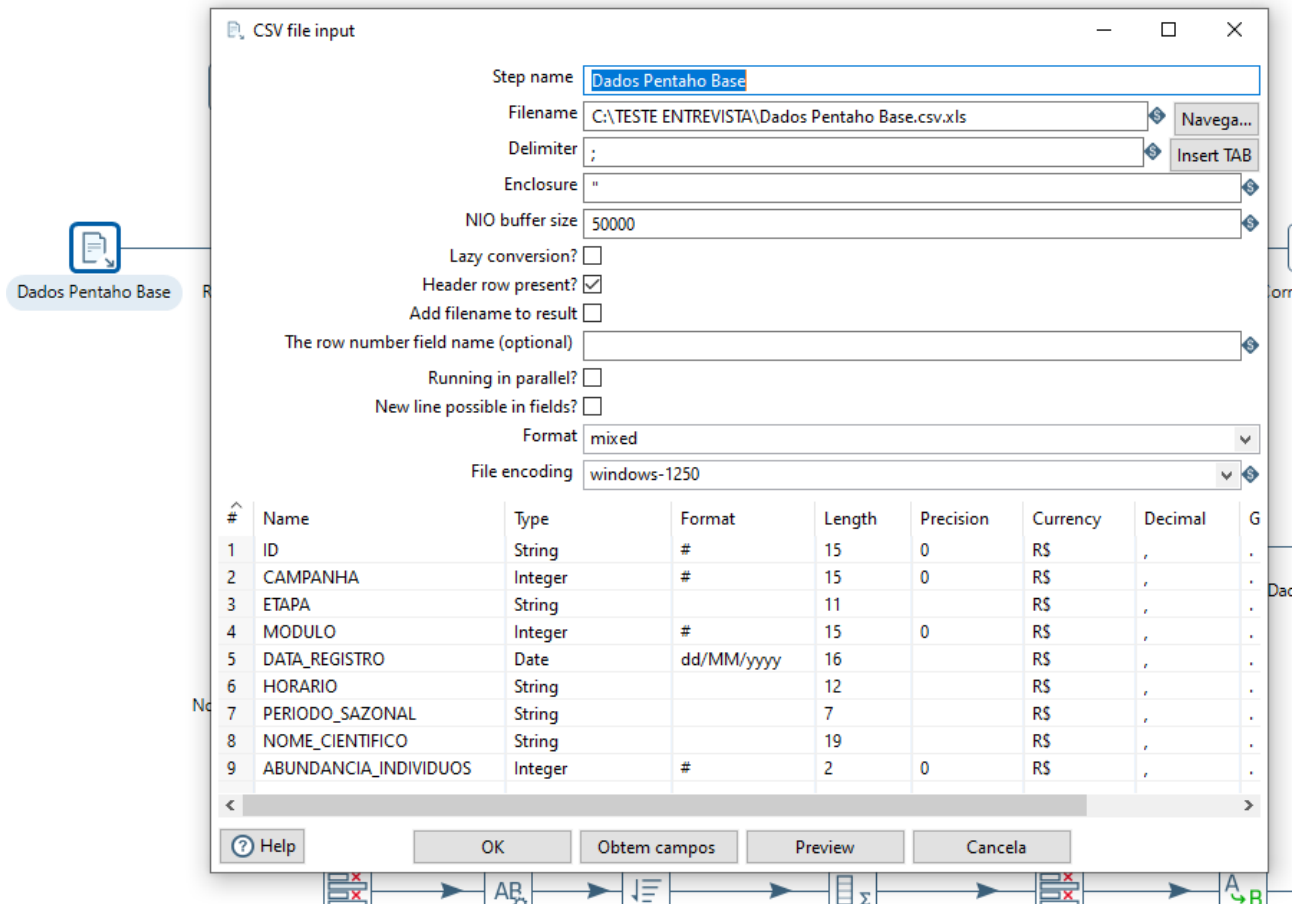
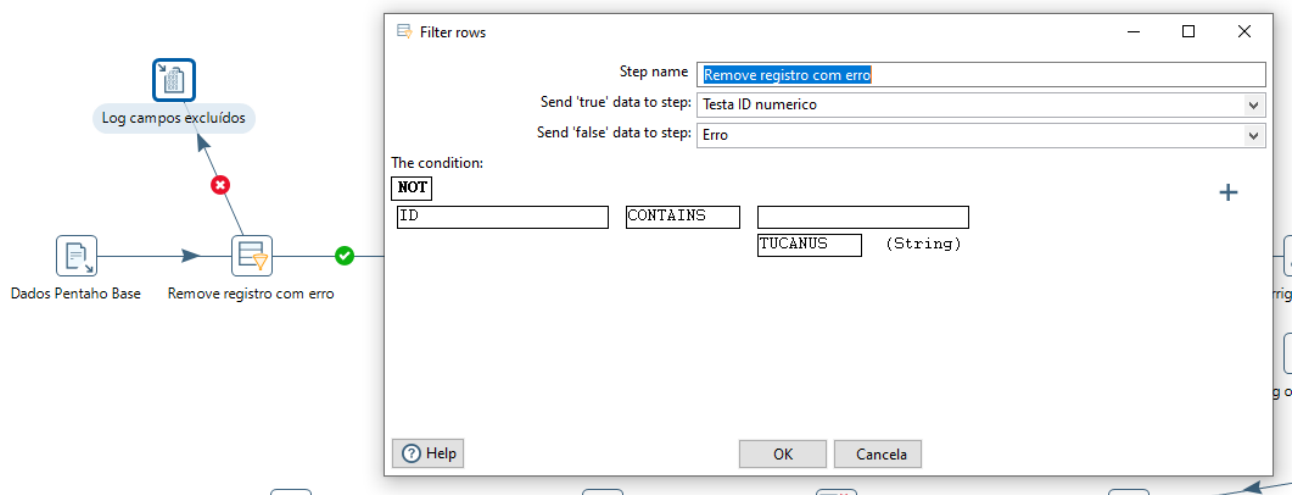


Pipeline de dados do monitoramento da abundância

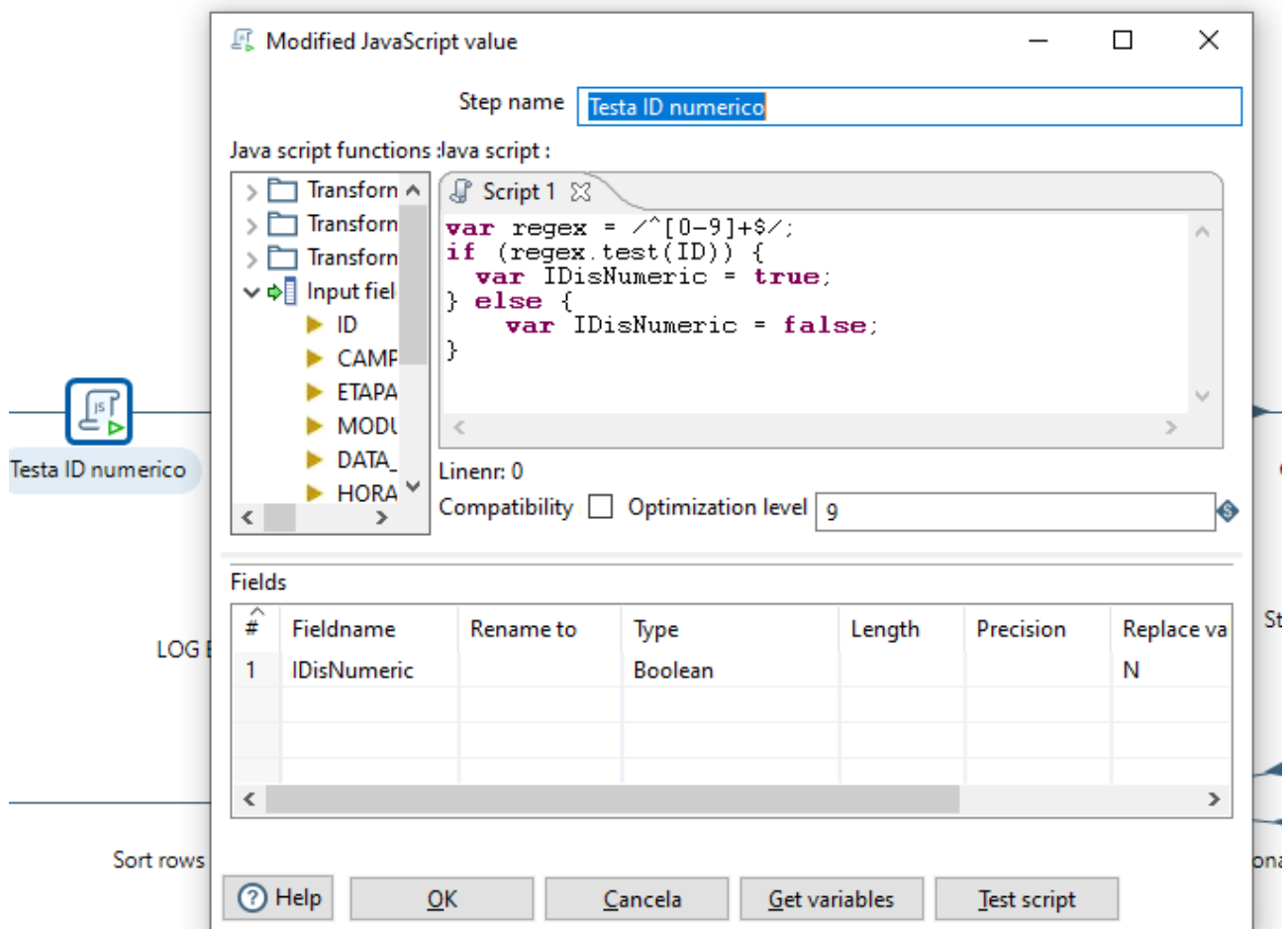
1. **Importação do csv de dados base;** foi feito já ajustando o campo data_registro para vir no formato correto; os campos ID e HORARIO, tiveram que ser importados como string, pois foi necessário efetuar limpezas, filtragem e padronização dos dados.



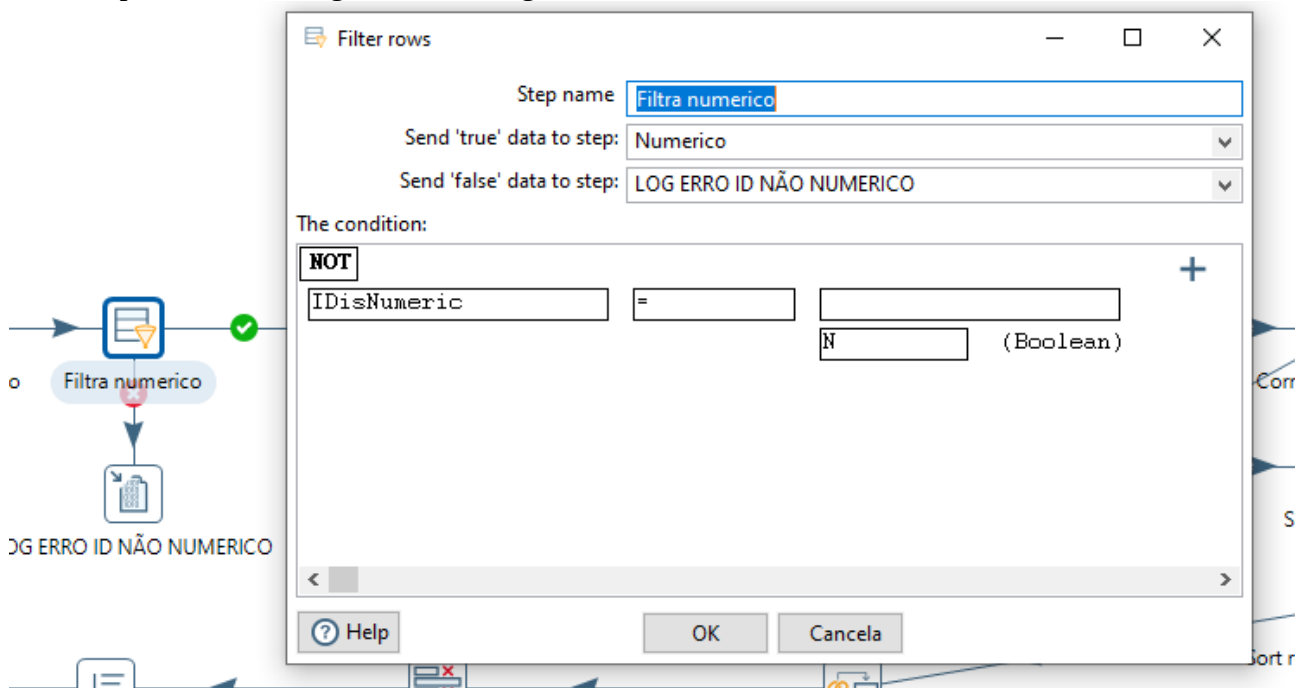
2. **Remove registros com erro (não numéricos);** no passo posterior apresentamos no log os registros não numéricos, para que desta forma possamos remover as linhas com erro, as linhas que forem excluídas serão apresentadas no log com a mensagem “REGISTRO EXCLUÍDO”.



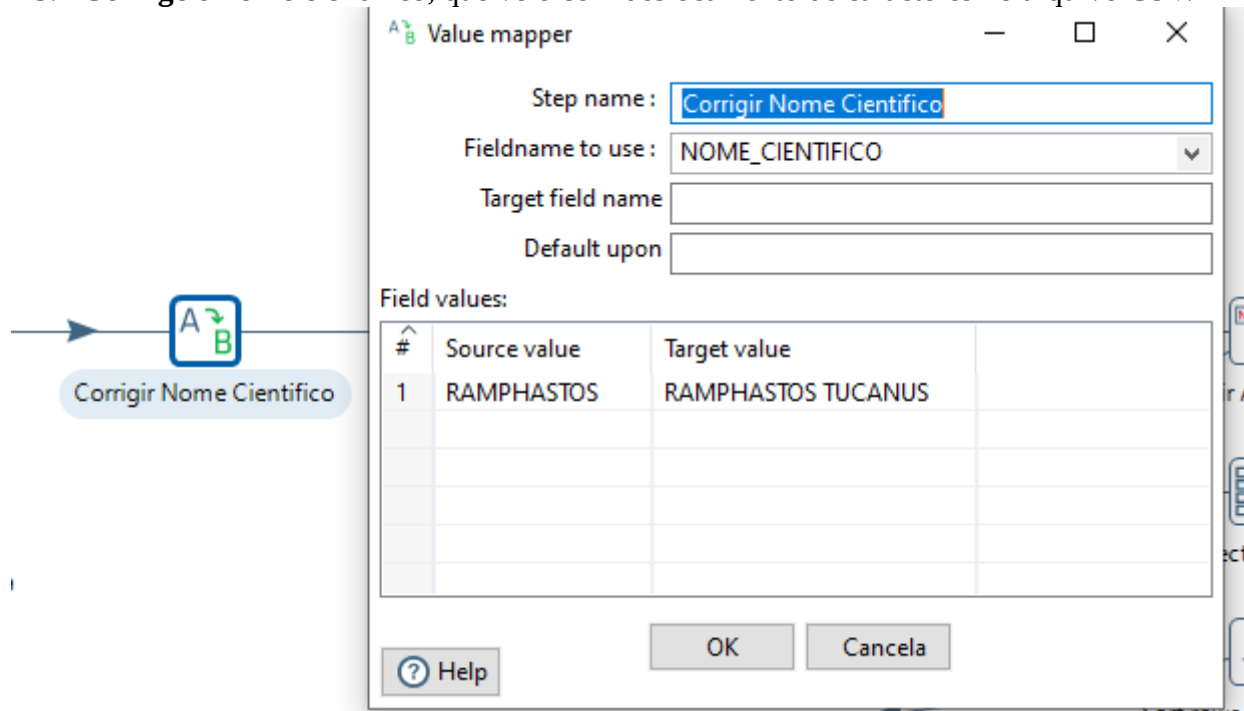
3. **Seleciona os IDs não numéricos;** usando javascript para fazer este teste.



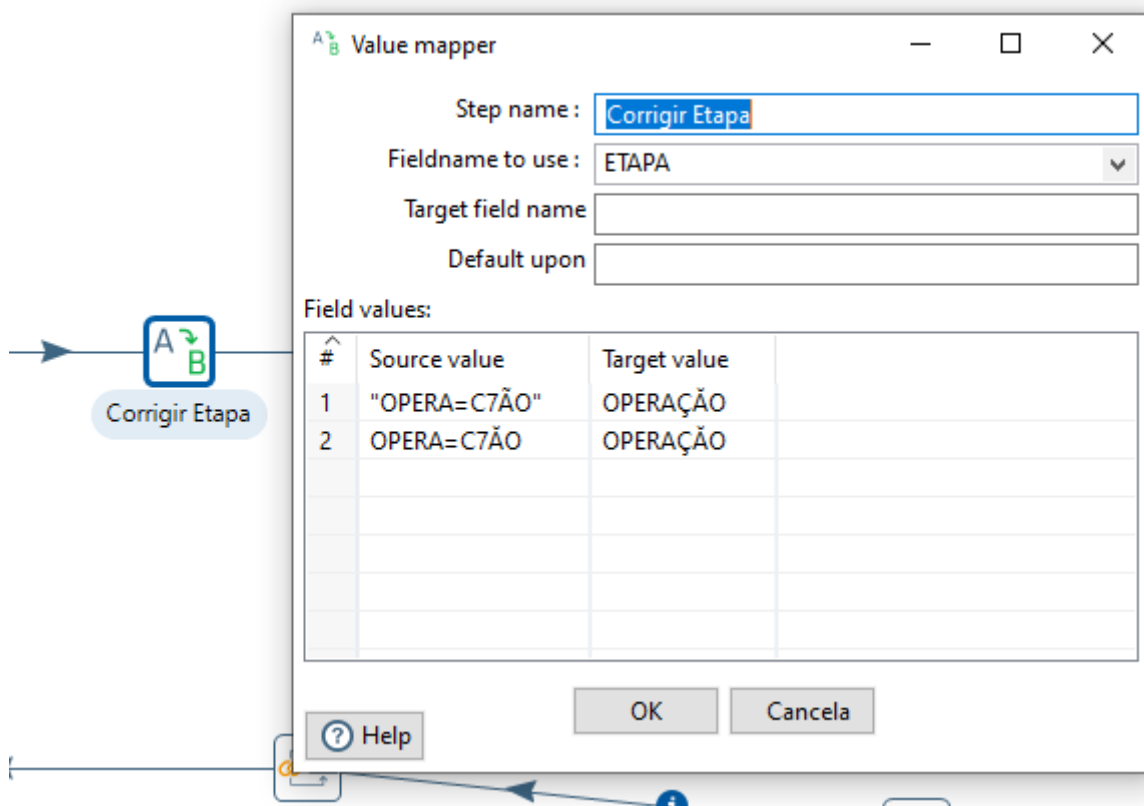
4. **Filtro para deixar passar somente os campos numéricos;** caso não seja numérico a linha é apresentada no log com a mensagem “ERRO CAMPO NÃO NUMERICO”.



5. **Corrige o nome científico**, que veio com deslocamento de caracteres no arquivo CSV.



6. **Corrige etapa que veio com o valor errado.** (OBS: este erro foi localizado nas análises no banco de dados)



7. **Corrige dados de horário**, para possibilitar junção a data registro; (OBS: o levantamento de tipos de erros foi feito no Excel)

Modified JavaScript value

Step name: **Corrigir Horário**

Java script functions:

- Transform Scripts
- Transform Constants
- Transform Functions
- Input fields
 - ID
 - CAMPANHA
 - ETAPA
 - MODULO
 - DATA_REGISTRO
 - HORARIO
 - PERIODO_SAZONA
 - NOME_CIENTIFICO
 - ABUNDANCIA_IND
 - IDisNumeric
- Output fields

Please use the 'Reg' button to add new fields.

Java script:

```
Script 1
if (!HORARIO){
  //Coloca o valor 99:99 para saber que o campo estava nulo
  var HORARIO_TRATADO = '00:00';
} else if (HORARIO.indexOf(':') != -1){
  //trata os dois tipos de campos '1900/01/01 09:40:00.000' e '20:50:00.000'
  var HORARIO_TRATADO = HORARIO.substr(HORARIO.indexOf(':') - 2, 5 );
} else if (HORARIO.indexOf('h') != -1){
  //trata os campos no formato '6h14'
  var HORARIO_TRATADO = '0'+ HORARIO.replace('h', ':');
} else if (HORARIO.indexOf(';') != -1){
  //trata os campos no formato '6:14'
  var HORARIO_TRATADO = '0'+ HORARIO.replace(':', ':');
}
```

Linens: 0

Compatibility mode? ☐ Optimization level: 9

Fields

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	HORARIO_TRATADO		String			N

Buttons: Help, OK, Cancela, Get variables, Test script

Workflow steps: Corrigir Horário, String operation, Dados Adicionais

8. **Corrige erro causado por deslocamento de caracteres.**

If field value is null

Step name: **Corrigir Abundancia**

Replace Null for all fields

Replace by value: [dropdown]

Set empty string? ☐

Mask (Date): [dropdown]

Select fields ☒

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)	Set empty string?
1	ABUNDANCIA_INDIVIDUOS	1		N

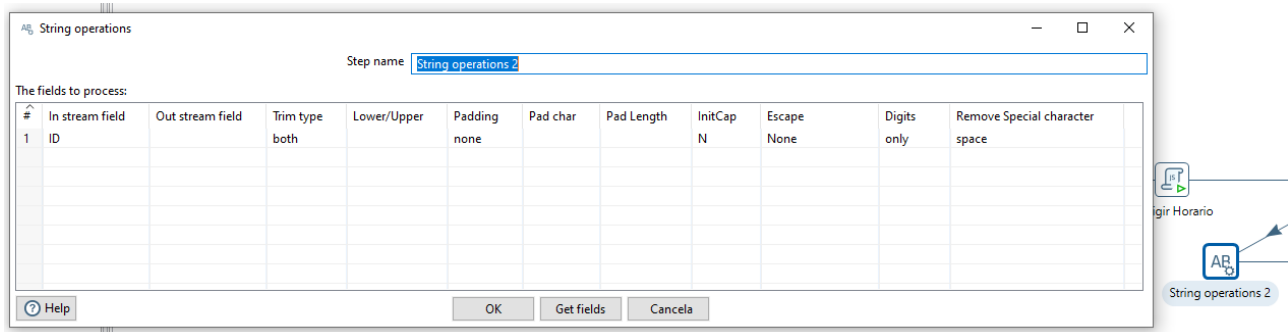
Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	ABUNDANCIA_INDIVIDUOS	1		N

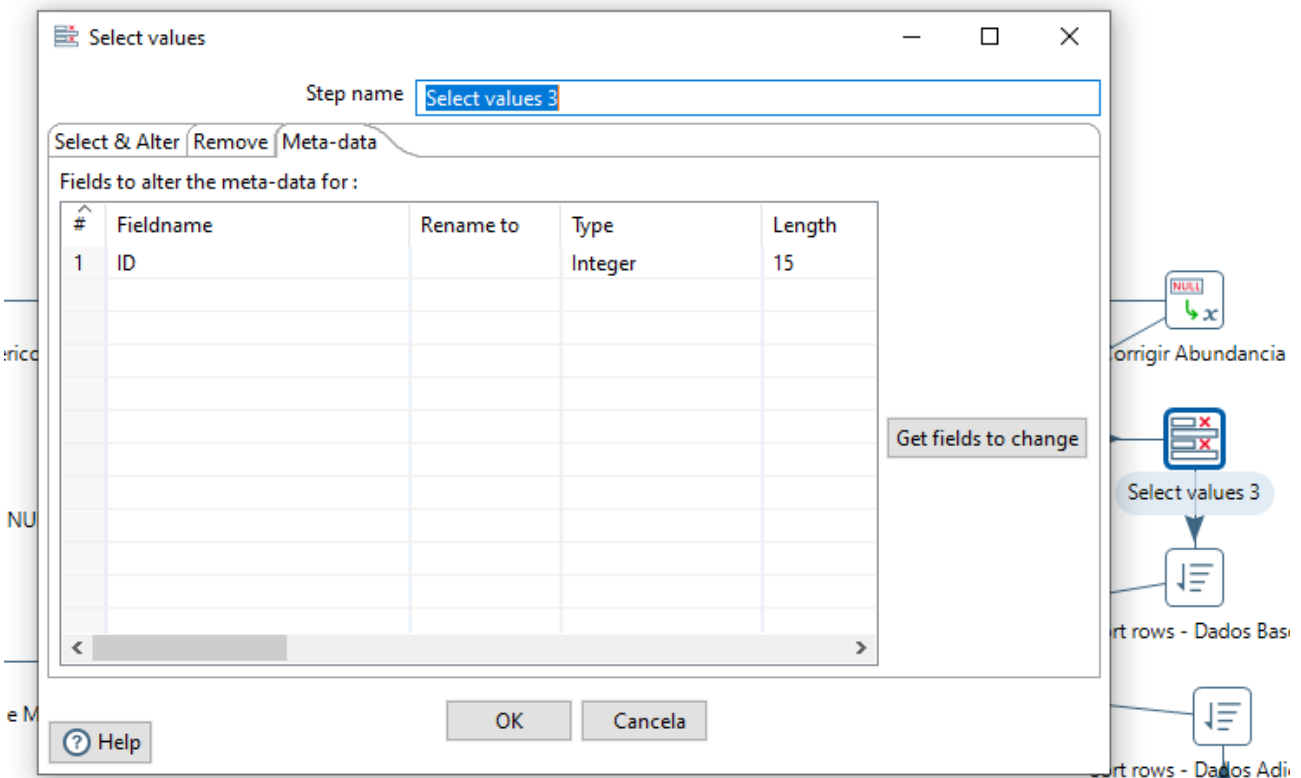
Buttons: Help, OK, Obtem campos, Cancela

Workflow steps: Corrigir Abundancia, Select values 3, Sort rows - Dados Base, Sort rows - Dados Adicionais, Dados Adicionais

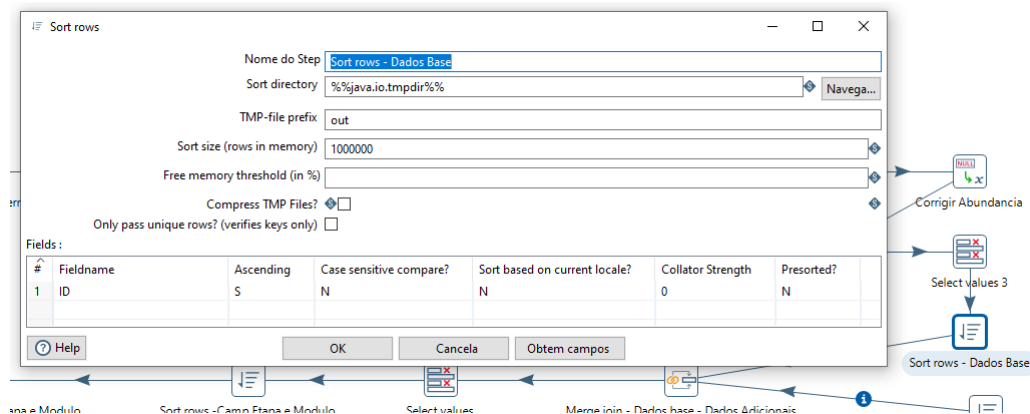
9. Remove espaços em branco e o caractere de space.



10. Faz o cast do campo ID de string para inteiro.



11. Faz a ordenação dos dados pelo campo ID, para possibilitar perfeita junção de dados com Dados adicionais.



12. Importa CSV de dados adicionais, já com os tipos corretos.

The 'CSV file input' dialog box is shown with the following settings:

- Step name: Dados Adicionais
- Filename: C:\TESTE ENTREVISTA\Dados Pentaho Adicional.csv.xls
- Delimiter: ;
- Enclosure: "
- NIO buffer size: 50000
- Lazy conversion? ☒
- Header row present? ☒
- Add filename to result? ☐
- The row number field name (optional):
- Running in parallel? ☐
- New line possible in fields? ☐
- Format: mixed
- File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim
1	ID	Integer	#	15	0	R\$,	.	nen
2	STATUS_EXTINCAO	String		9		R\$,	.	nen

The diagram shows the following steps in a data flow:

- Corrigir Abundancia
- Select values 3
- Sort rows - Dados Base
- Sort rows - Dados Adicionais
- Dados Adicionais
- Corrigir Status

Buttons at the bottom of the dialog: Help, OK, Obtem campos, Preview, Cancela.

13. Merge feito para junção das duas informações.

The 'Merge join' dialog box is shown with the following settings:

- Step name: Merge join - Dados base - Dados Adicionais
- First Step: Sort rows - Dados Base
- Second Step: Sort rows - Dados Adicionais
- Join Type: INNER

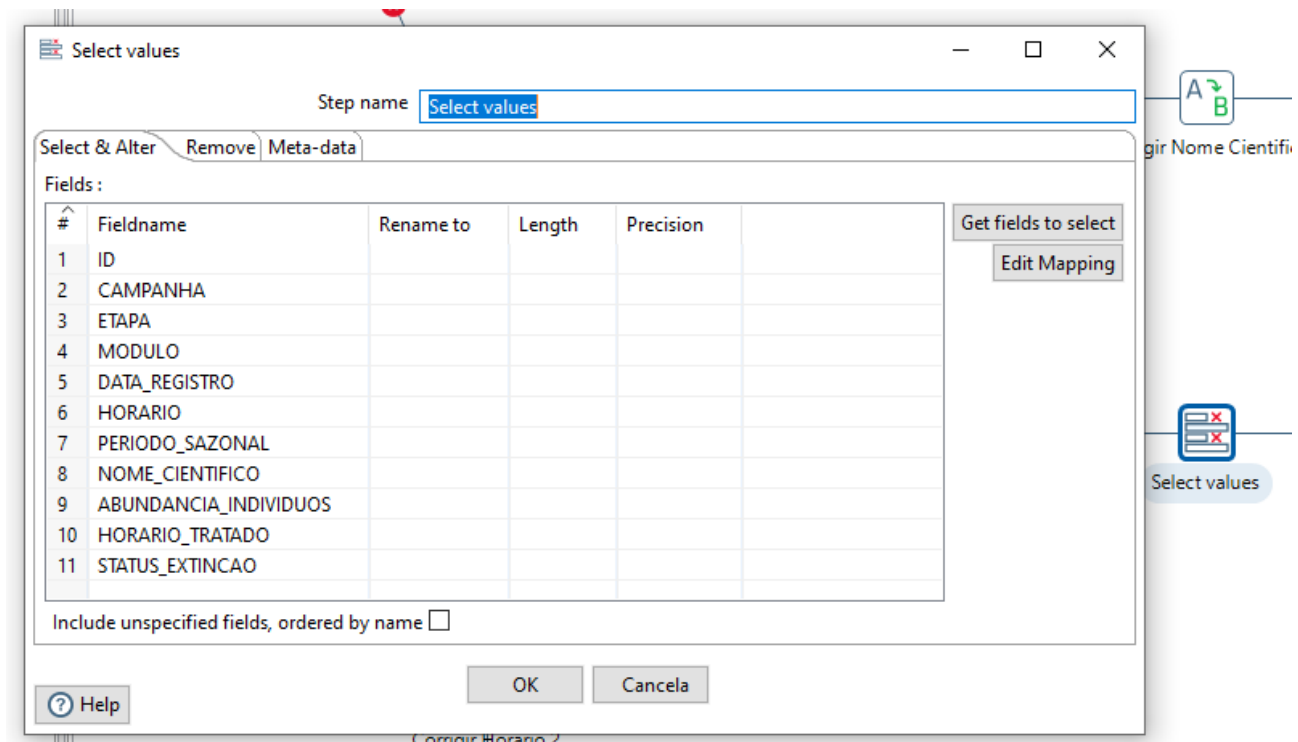
Keys for 1st step:		Keys for 2nd step:	
#	Key field	#	Key field
1	ID	1	ID

Buttons at the bottom of the dialog: Help, OK, Cancela.

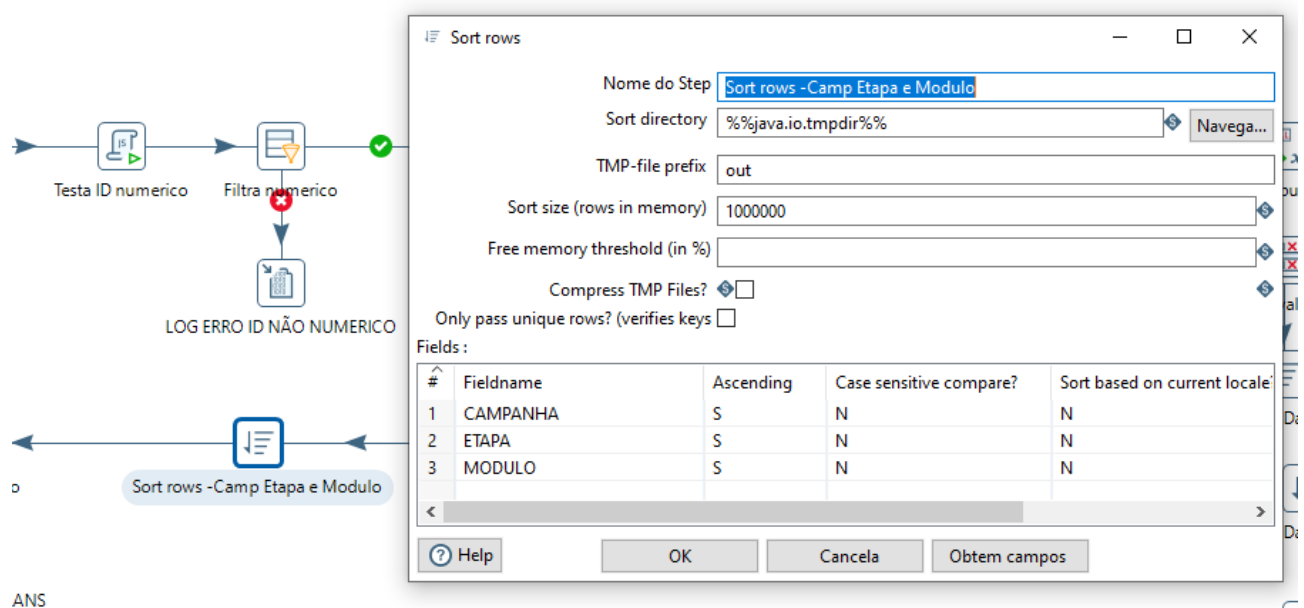
The diagram shows the following steps in a data flow:

- String c
- Merge join - Dados base - Dados Adicionais

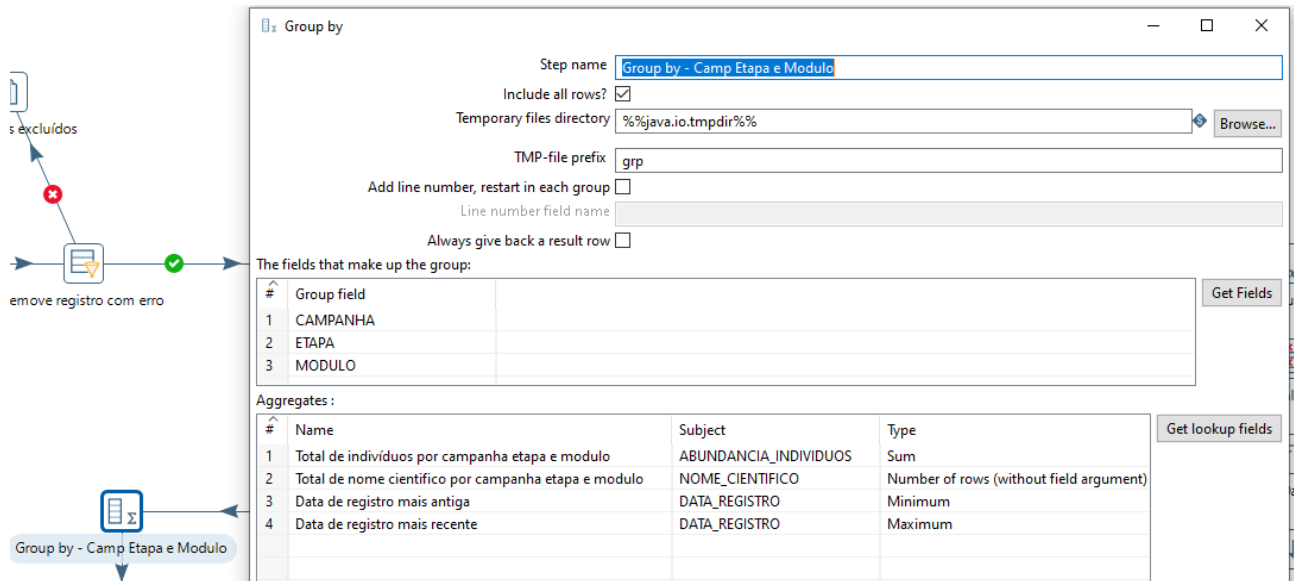
14. Seleção dos campos necessários.



15. Ordenação para possibilitar o group by.



16. **Group by por campanha, etapa e modulo**, para possibilitar a soma, o count, min e max.



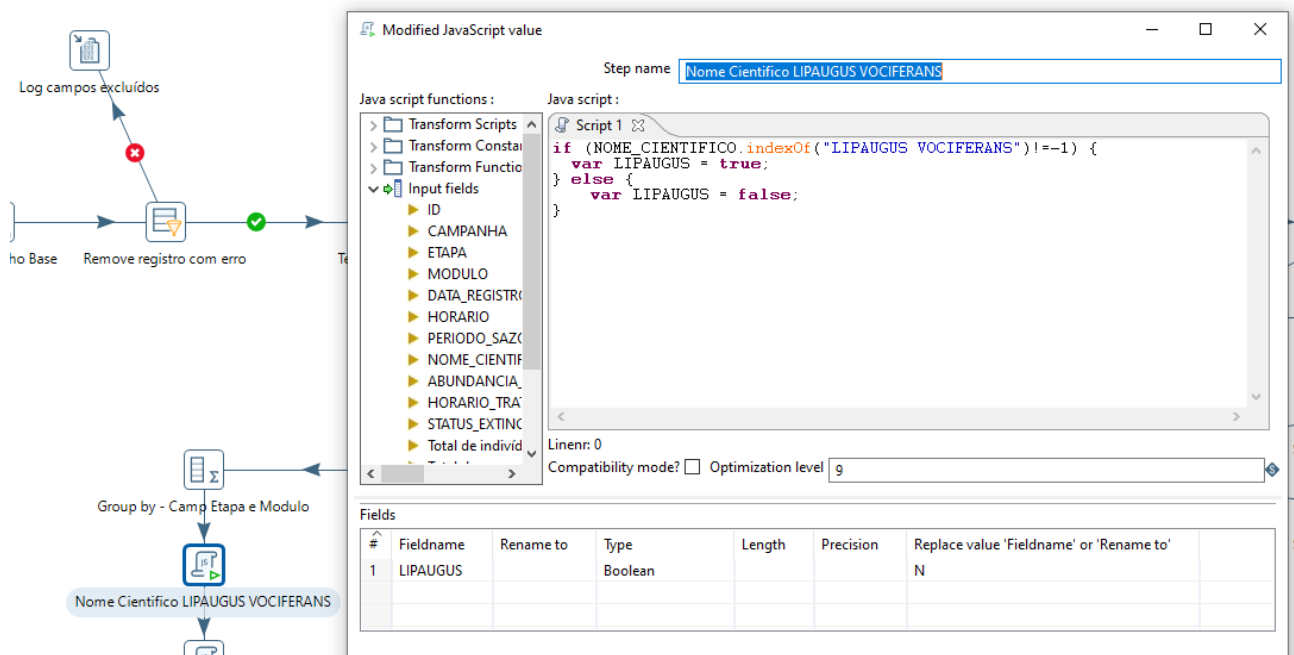
The screenshot shows the 'Group by' configuration window. The step name is 'Group by - Camp Etapa e Modulo'. The 'Include all rows?' checkbox is checked. The 'Temporary files directory' is set to '%java.io.tmpdir%'. The 'TMP-file prefix' is 'grp'. The 'Add line number, restart in each group' checkbox is unchecked. The 'Line number field name' is empty. The 'Always give back a result row' checkbox is unchecked. The 'The fields that make up the group:' table lists the following fields:

#	Group field
1	CAMPANHA
2	ETAPA
3	MODULO

The 'Aggregates:' table lists the following aggregates:

#	Name	Subject	Type
1	Total de individuos por campanha etapa e modulo	ABUNDANCIA_INDIVDUOS	Sum
2	Total de nome cientifico por campanha etapa e modulo	NOME_CIENTIFICO	Number of rows (without field argument)
3	Data de registro mais antiga	DATA_REGISTRO	Minimum
4	Data de registro mais recente	DATA_REGISTRO	Maximum

17. **JavaScript que verifica se o nome é "LIPAUGUS VOCIFERANS"**, e cria um campo booleano.



The screenshot shows the 'Modified JavaScript value' configuration window. The step name is 'Nome Cientifico LIPAUGUS VOCIFERANS'. The 'JavaScript functions' list shows 'Input fields' expanded. The 'JavaScript' code block contains the following script:

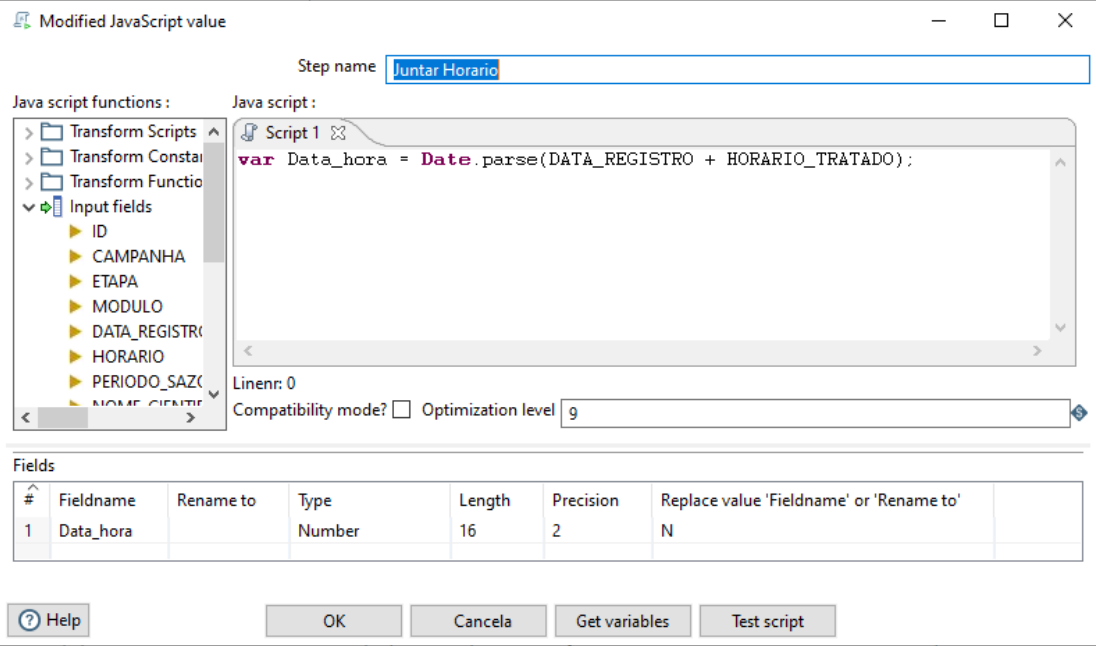
```

if (NOME_CIENTIFICO.indexOf("LIPAUGUS VOCIFERANS") != -1) {
    var LIPAUGUS = true;
} else {
    var LIPAUGUS = false;
}
    
```

The 'Fields' table lists the following fields:

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	LIPAUGUS		Boolean			N

18. **JavaScript para junção da hora com a data**, (OBS: com o step concat fields apresentou erro de null pointer, com o java script solucionou o erro)



registro com erro

by - Camp Etapa e

ntifico LIPAUGUS VOCIF

Juntar Horário

Altera tipo Data_hora

Modified JavaScript value

Step name: Juntar Horário

JavaScript functions:

- Transform Scripts
- Transform Constant
- Transform Function
- Input fields
 - ID
 - CAMPANHA
 - ETAPA
 - MODULO
 - DATA_REGISTRO
 - HORARIO
 - PERIODO_SAZO
 - ...

JavaScript:

```
var Data_hora = Date.parse(DATA_REGISTRO + HORARIO_TRATADO);
```

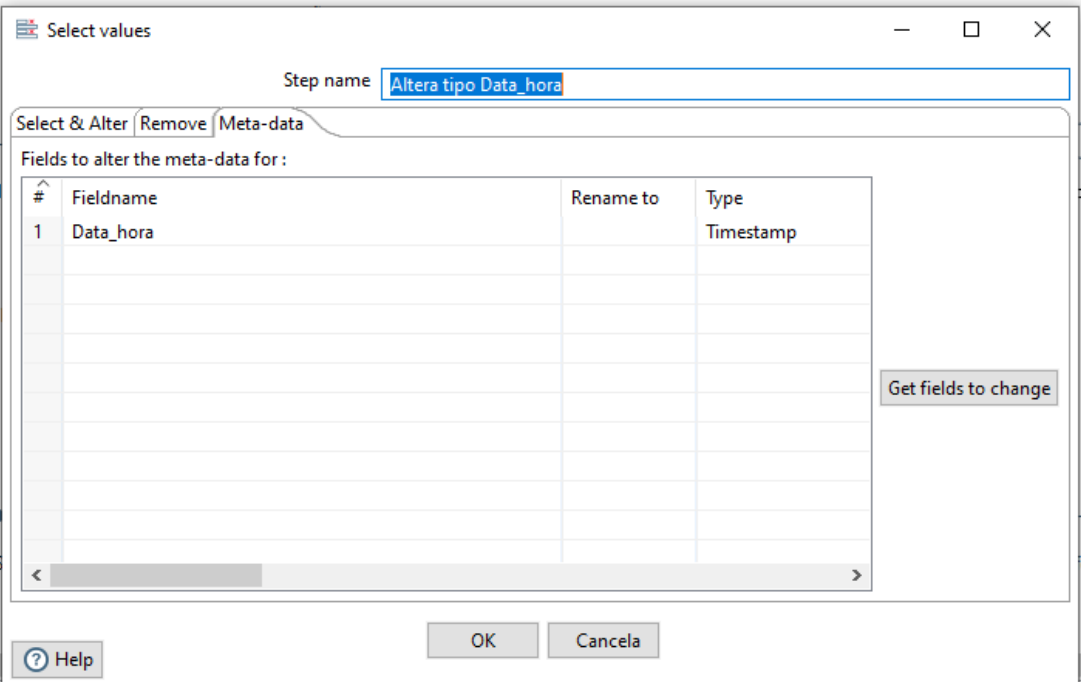
Linenn: 0

Compatibility mode? ☐ Optimization level 9

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	Data_hora		Number	16	2	N

Buttons: Help, OK, Cancela, Get variables, Test script

19. **Faz o cast no campo data_hora para Timestamp.**



by - Camp Etapa e Mod

ntifico LIPAUGUS VOCIF

Juntar Horário

Altera tipo Data_hora

Select values

Step name: Altera tipo Data_hora

Select & Alter Remove Meta-data

Fields to alter the meta-data for:

#	Fieldname	Rename to	Type
1	Data_hora		Timestamp

Get fields to change

Buttons: Help, OK, Cancela

20. Tira os espaços vazios (trim) e coloca o campo “PERIODO SAZONAL” para caixa baixa.

String operations

Step name: String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	PERIODO_SAZONAL		both	lower	none			N	None	none	none
2	NOME_CIENTIFICO		both								
3	HORARIO		both								

OK Get fields Cancela

Altera tipo Data_hora String operations Sort rows Group by - data_registro Qtd dia distintos Excluir mascara Get system info

21. Ordena baseado na DATA_REGISTRO, para possibilitar o group by.

Sort rows

Nome do Step: Sort rows

Sort directory: %%java.io.tmpdir%%

TMP-file prefix: out

Sort size (rows in memory): 1000000

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?
1	DATA_REGISTRO	S	N	N

OK Cancela Obtem campos

Altera tipo Data_hora String operations Sort rows Group by - data_registro Qtd dia distintos Excluir mascara

22. Faz o group by pela DATA_REGISTRO e conta com distinct.

Group by

Step name: Group by - data_registro

Include all rows? ☒

Temporary files directory: %%java.io.tmpdir%% Browse...

TMP-file prefix: grp

Add line number, restart in each group ☐

Line number field name: teste

Always give back a result row ☐

The fields that make up the group:

#	Group field
1	DATA_REGISTRO

Get Fields

Aggregates:

#	Name	Subject	Type	Value
1	Qtd dias distintos	DATA_REGISTRO	Number of Distinct Values (N)	

Get lookup fields

Help OK Cancela

Workflow: Altera tipo Data_hora → String operations → Sort rows → **Group by - data_registro** → Qtd dia distintos → Excluir mascara → Get system info

23. Converte Qtd dias distintos para string para pode fazer o replace da mascara que ficou retornando no campo.

Replace in string

Step name: Excluir mascara

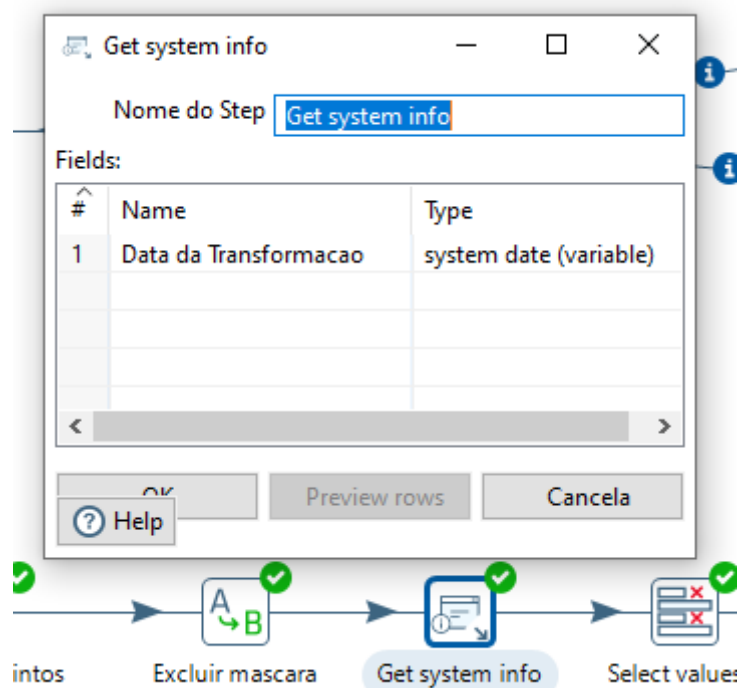
Fields string:

#	In stream field	Out stream field	use RegEx	Search	Replace with
1	Qtd dias distintos		N	dd/MM/yyyy	

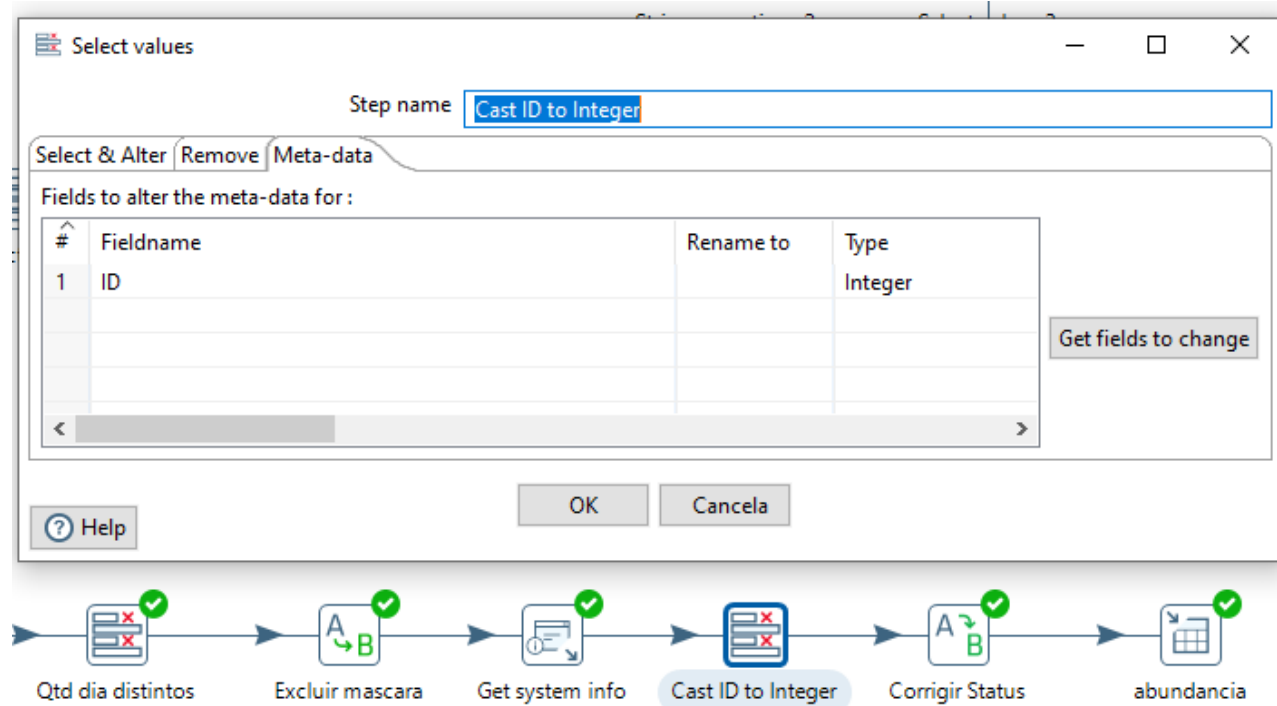
Help OK Get fields Cancela

Workflow: Sort rows → Group by - data_registro → Qtd dia distintos → **Excluir mascara** → Get system info → Select values 2

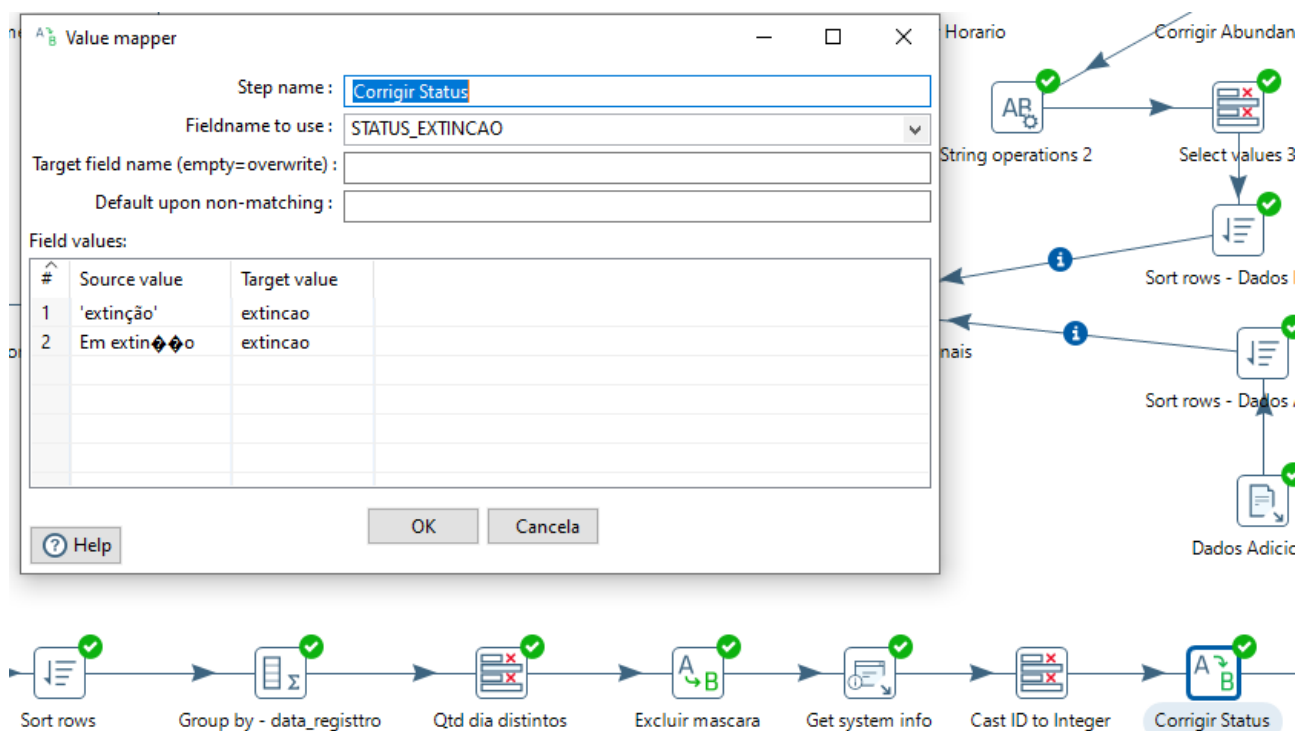
24. Recupero a data para poder salvar no campo de data da transformação.



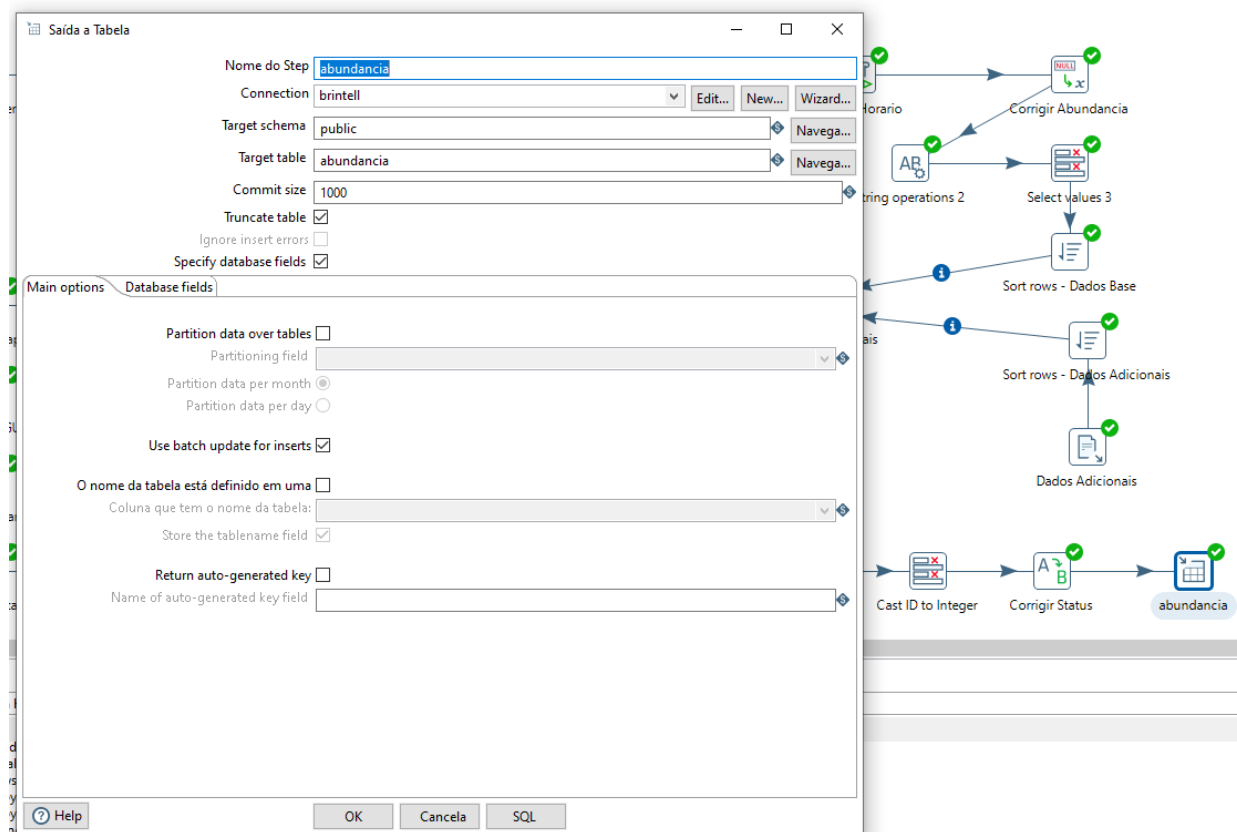
25. Cast do ID para Integer para possibilitar ser a PK.



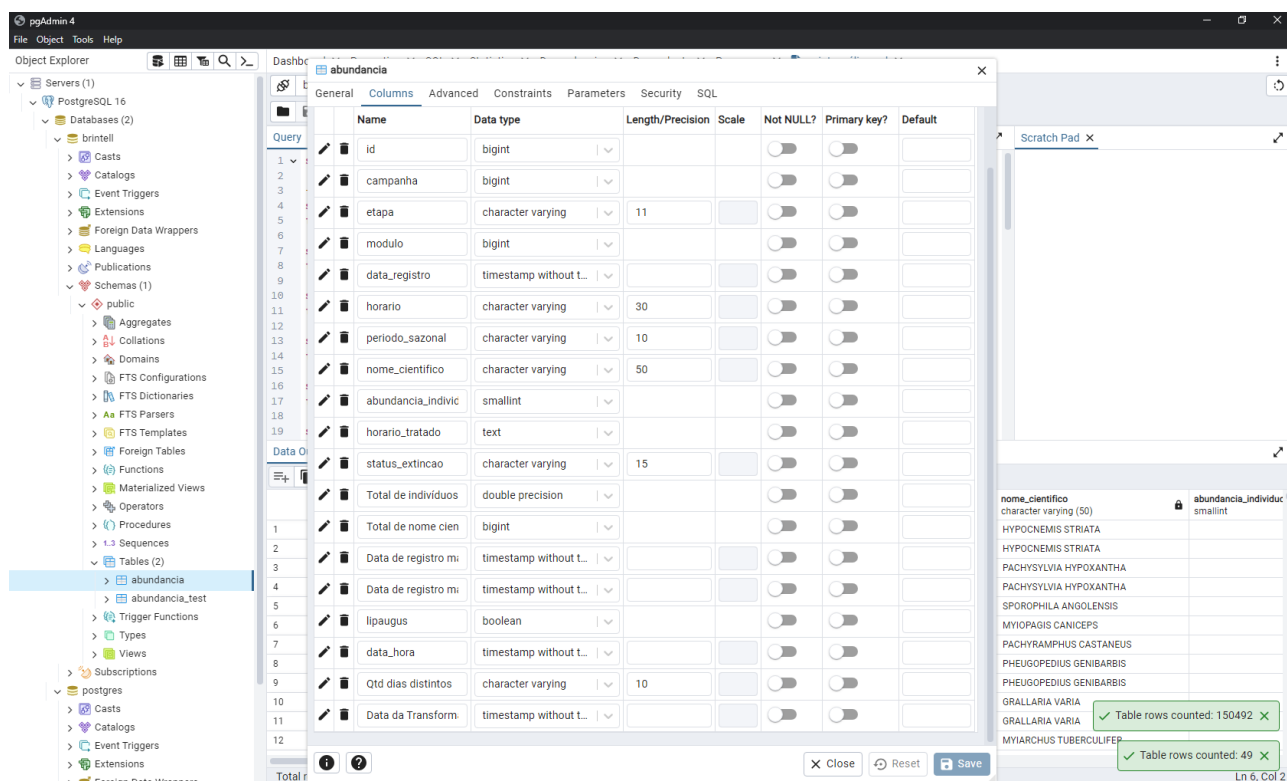
26. Remoção dos acentos do campo STATUS_EXTINCAO.



27. Persistência dos dados na tabela abundância na database brintell, no banco de dados Postgres.



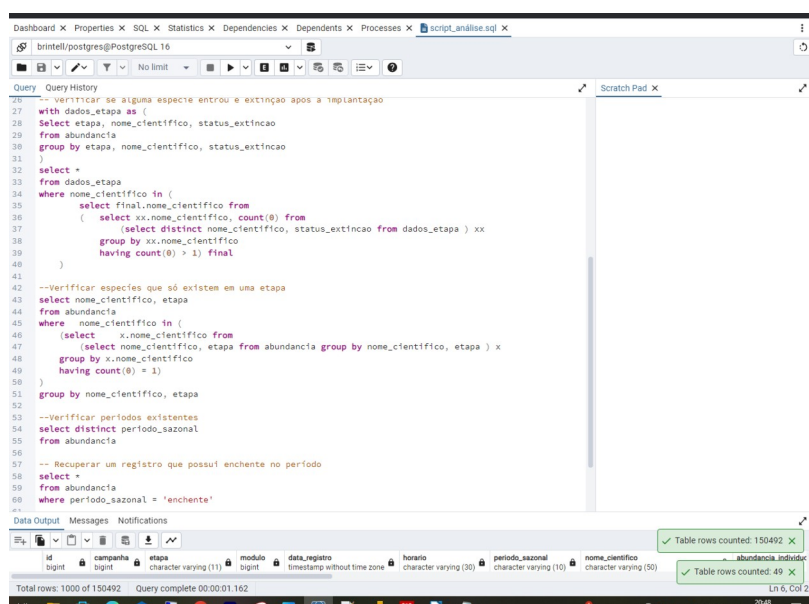
28. Base de dados e tabela criadas.



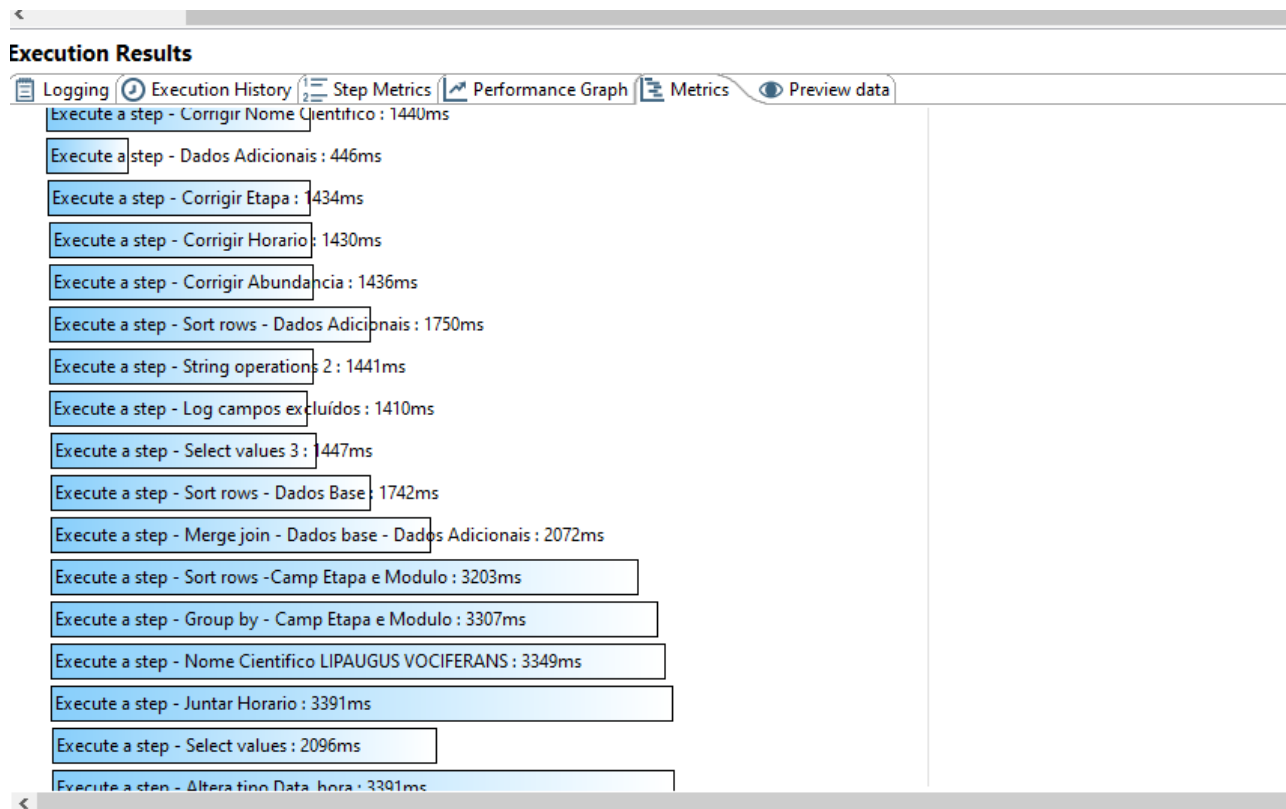
29. Consultas criadas para validação de dados e consistência. (Segue arquivo SQL), também levantei as espécies que não existiam durante a implantação e que existiam durante a operação e vice e versa.

OBS:

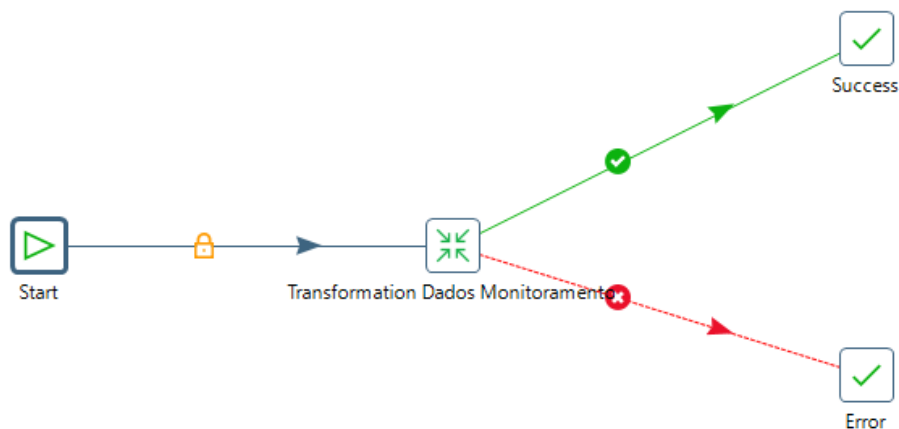
Nesta análise surgiu uma dúvida se os registros são incluídos um a um, por exemplo: Foram avistados 2 animais, neste caso teria dois registros com a mesma data e hora, ou se estes registros seriam duplicidades; se fossem seria fácil a limpeza adicionando uma step Unique rows, mas como não possuía um documento com regras decidi deixar os dados.

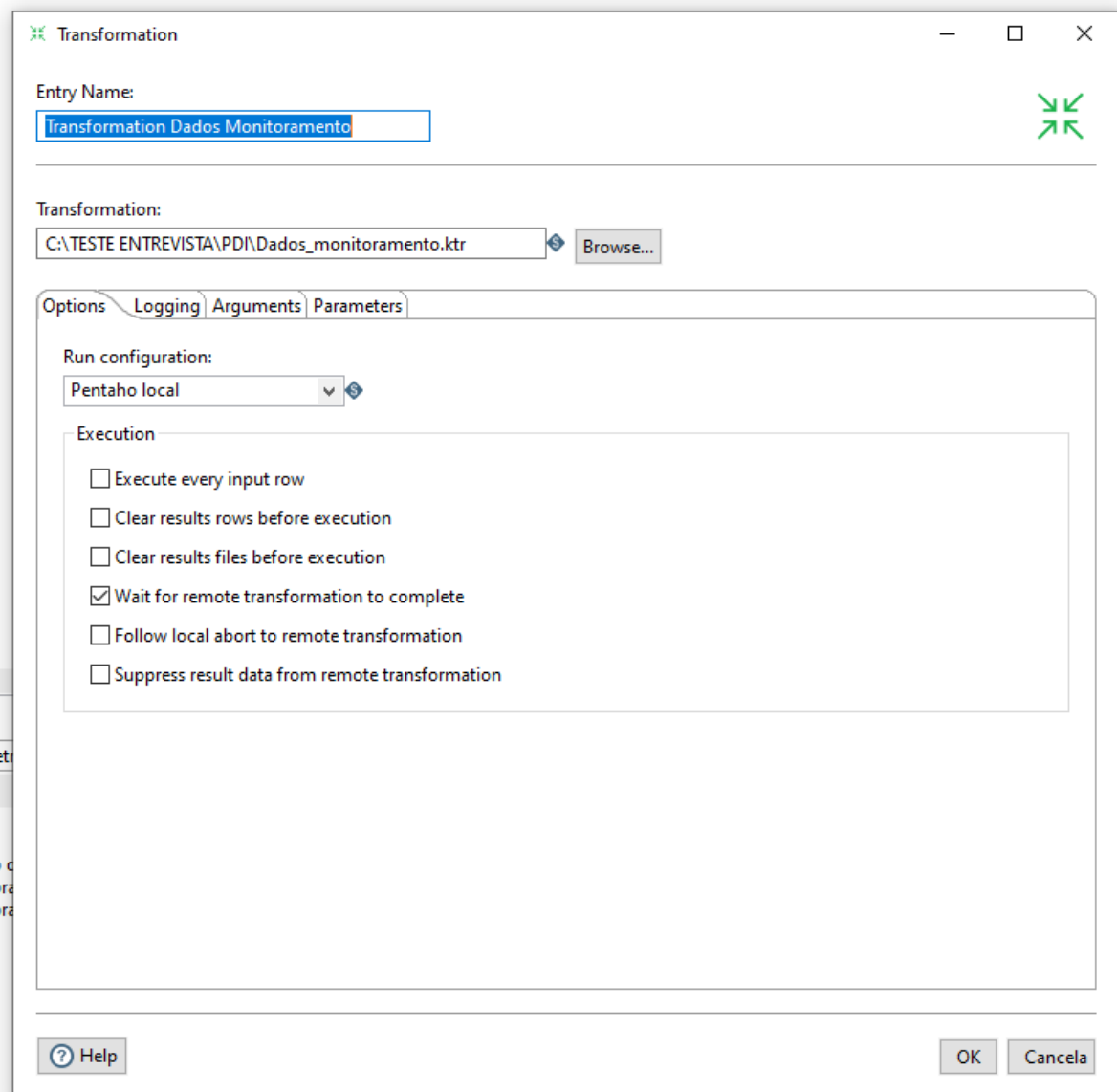


30. **Avaliando as métricas** pude ver que para ganho de performance seria considerável criar uma nova transformação para fazer os tratamentos no arquivo antes desta transformação, adicionando a transformação antes desta no Job ou em uma step Mapping(sub-transformation) no começo da transformação.

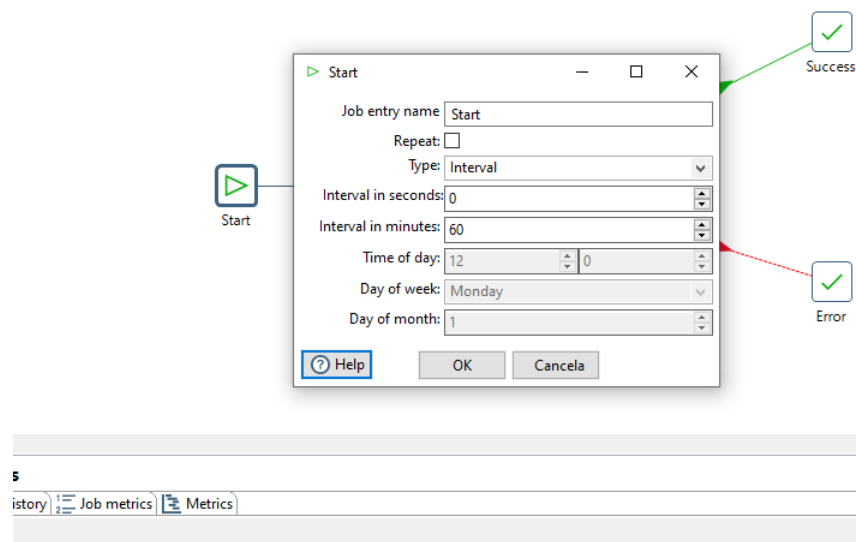


31. Job para agendamento criado





32. O Agendamento pode ser feito no próprio start.



33. Agendamento feito com arquivos .bat no schedule do próprio Sistema Operacional

```
@echo off

TITLE ExecutaJob
SET currentdir=%~dp0
SET kitchen=C:\Program Files\pentaho\data-integration\Kitchen.bat
SET logfile="%currentdir%log.txt"

echo. >> %logfile%
echo. >> %logfile%
"%kitchen%" /file:"C:\TESTE ENTREVISTA\PDI\Job_dado_monitoramento.kjb" /level:Basic >> %logfile%

call "C:\TESTE ENTREVISTA\PDI\executa_job_abundancia.bat"
```

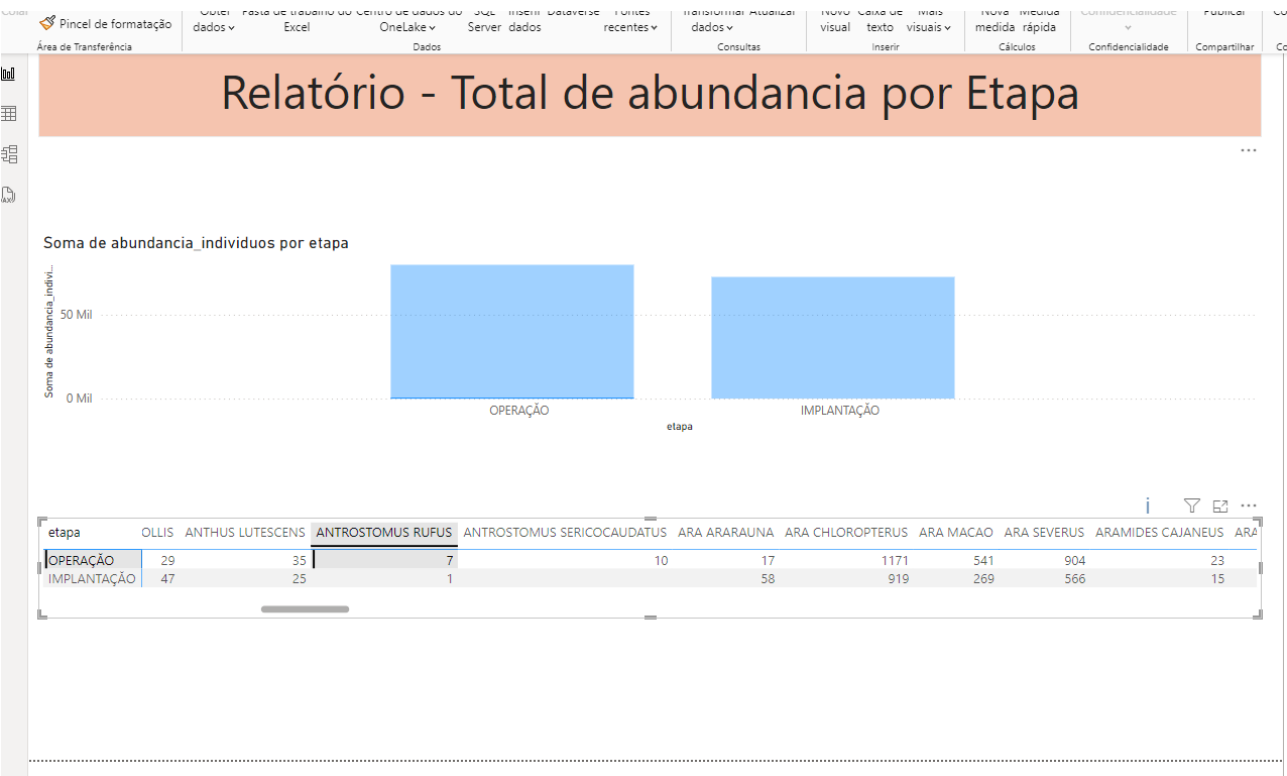
Execução com sucesso:

log.txt - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

2024/05/23 22:06:32 - Nome Cientifico LIPAUGUS VOCIFERANS.0 - Optimization level set to 9.
2024/05/23 22:06:32 - Juntar Horario.0 - Optimization level set to 9.
2024/05/23 22:06:33 - Group by - Camp Etapa e Modulo.0 - Linenr 50000
2024/05/23 22:06:33 - Nome Cientifico LIPAUGUS VOCIFERANS.0 - linenr 50000
2024/05/23 22:06:33 - Juntar Horario.0 - linenr 50000
2024/05/23 22:06:33 - Altera tipo Data_hora.0 - linenr 50000
2024/05/23 22:06:33 - Sort rows.0 - Linenr 50000
2024/05/23 22:06:34 - Group by - Camp Etapa e Modulo.0 - Linenr 100000
2024/05/23 22:06:34 - Nome Cientifico LIPAUGUS VOCIFERANS.0 - linenr 100000
2024/05/23 22:06:34 - Juntar Horario.0 - linenr 100000
2024/05/23 22:06:34 - Altera tipo Data_hora.0 - linenr 100000
2024/05/23 22:06:34 - Sort rows.0 - Linenr 100000
2024/05/23 22:06:34 - Sort rows -Camp Etapa e Modulo.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:34 - Group by - Camp Etapa e Modulo.0 - Linenr 150000
2024/05/23 22:06:34 - Group by - Camp Etapa e Modulo.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:34 - Nome Cientifico LIPAUGUS VOCIFERANS.0 - linenr 150000
2024/05/23 22:06:34 - Nome Cientifico LIPAUGUS VOCIFERANS.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:34 - Juntar Horario.0 - linenr 150000
2024/05/23 22:06:34 - Altera tipo Data_hora.0 - linenr 150000
2024/05/23 22:06:34 - Sort rows.0 - Linenr 150000
2024/05/23 22:06:34 - Juntar Horario.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:34 - Altera tipo Data_hora.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:34 - String operations.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:36 - Group by - data_registtro.0 - Linenr 50000
2024/05/23 22:06:36 - Qtd dia distintos.0 - linenr 50000
2024/05/23 22:06:38 - Cast ID to Integer.0 - linenr 50000
2024/05/23 22:06:38 - Group by - data_registtro.0 - Linenr 100000
2024/05/23 22:06:39 - abundancia.0 - linenr 50000
2024/05/23 22:06:39 - Qtd dia distintos.0 - linenr 100000
2024/05/23 22:06:40 - Cast ID to Integer.0 - linenr 100000
2024/05/23 22:06:41 - Sort rows.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:41 - Group by - data_registtro.0 - Linenr 150000
2024/05/23 22:06:41 - Group by - data_registtro.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:42 - abundancia.0 - linenr 100000
2024/05/23 22:06:42 - Qtd dia distintos.0 - linenr 150000
2024/05/23 22:06:42 - Qtd dia distintos.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:42 - Excluir mascara.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:43 - Get system info.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:43 - Cast ID to Integer.0 - linenr 150000
2024/05/23 22:06:43 - Cast ID to Integer.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:44 - Corrigir Status.0 - Finished processing (I=0, O=0, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:44 - abundancia.0 - linenr 150000
2024/05/23 22:06:44 - abundancia.0 - Finished processing (I=0, O=150492, R=150492, W=150492, U=0, E=0)
2024/05/23 22:06:44 - Job_dado_monitoramento - Starting entry [Success]
2024/05/23 22:06:44 - Job_dado_monitoramento - Finished job entry [Success] (result=[true])
2024/05/23 22:06:44 - Job_dado_monitoramento - Finished job entry [Transformation Dados Monitoramento] (result=[true])
2024/05/23 22:06:44 - Job_dado_monitoramento - Job execution finished
2024/05/23 22:06:44 - Kitchen - Finished!
2024/05/23 22:06:44 - Kitchen - Start=2024/05/23 22:06:26.801, Stop=2024/05/23 22:06:44.759
2024/05/23 22:06:44 - Kitchen - Processing ended after 17 seconds.

34. Neste relatório do PowerBI; podemos notar que na implantação possui menos espécies que na operação, e logo abaixo que existem espécies na operação e não existem na implantação, podendo comprar seus números.



Geraldo José Ferreira Neto