FIAP

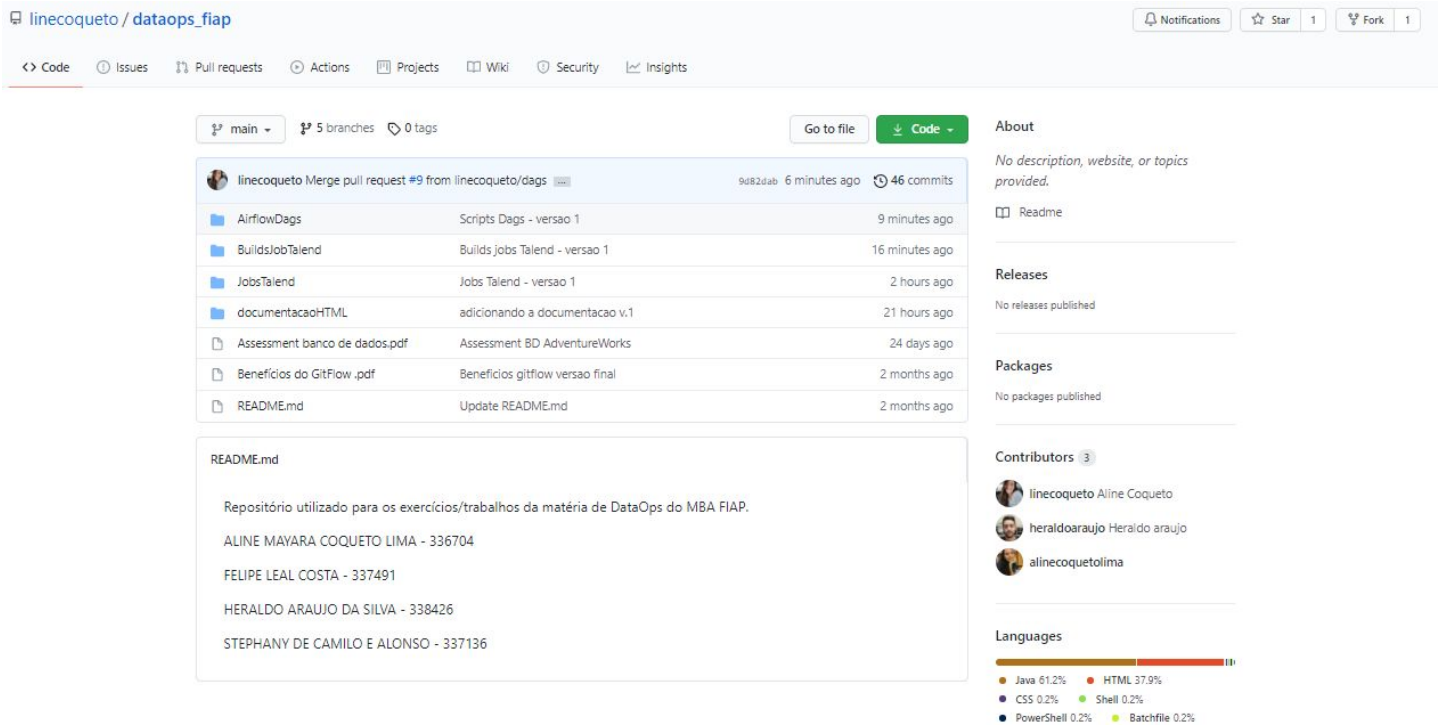
MBA+

Engenharia de Dados
DataOps

Aline	336704
Felipe	337491
Heraldo	338426
Stephany	337136

Projeto DataOps

O projeto descrito neste documento foi desenvolvido utilizando as ferramentas Talend e Airflow e publicado no GitHub através do GitFlow. Pode ser acessado em:
https://github.com/linecoqueto/dataops_fiap

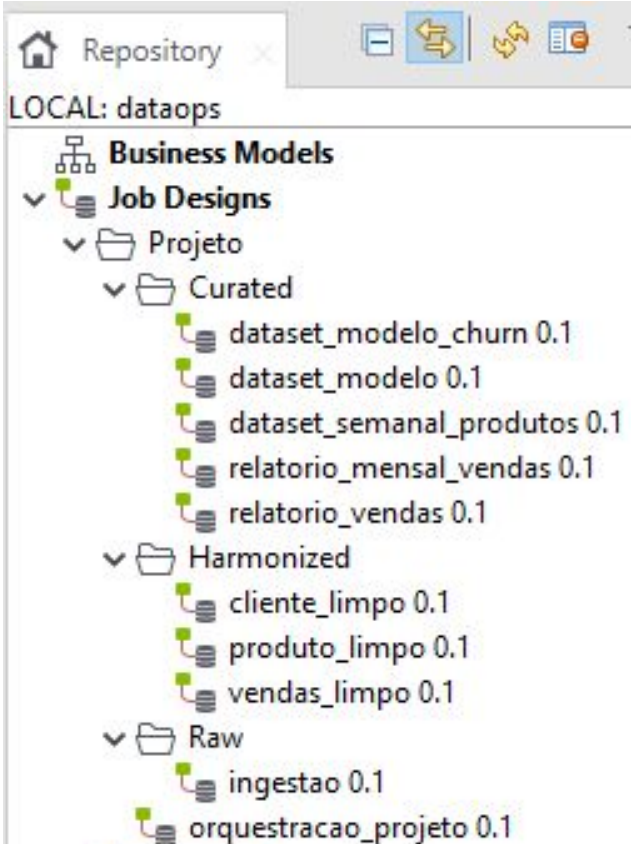


Arquitetura de Dados

O processamento dos dados foi realizado baseando a Arquitetura de Dados em três camadas: Raw, Harmonized e Curated. Tais camadas possuem as características descritas na figura abaixo.

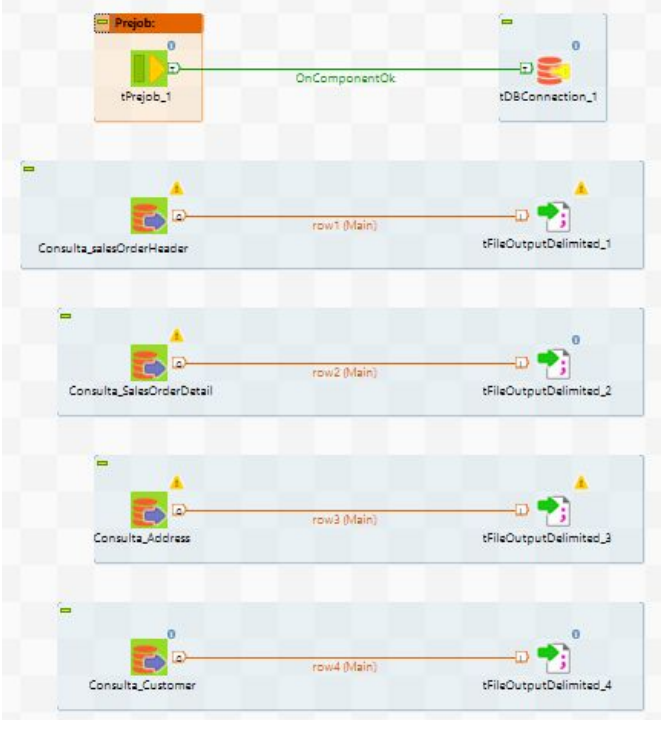
Arquitetura de Dados

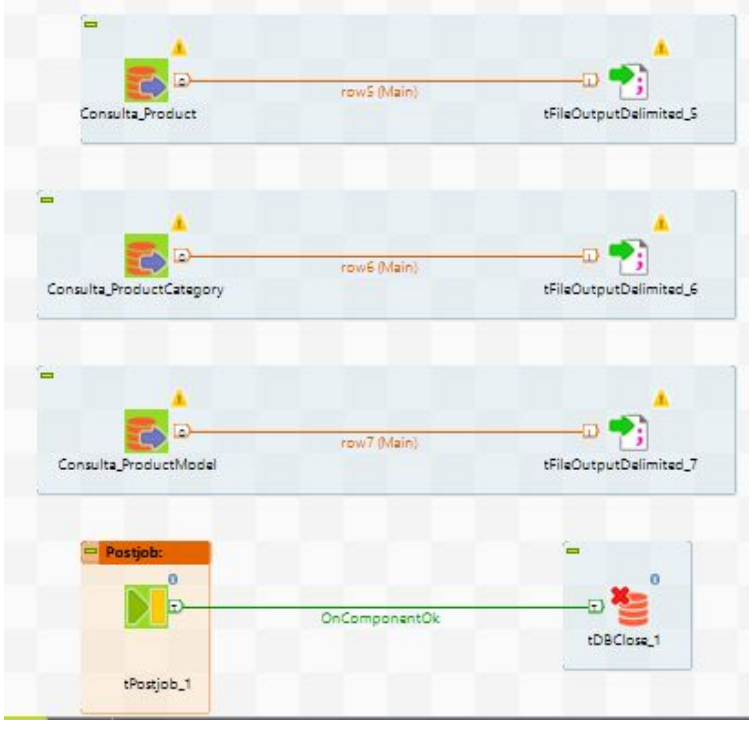
Na figura ao lado é mostrada a arquitetura de camadas construída através dos jobs no Talend. Nos próximos slides, cada job será detalhado.



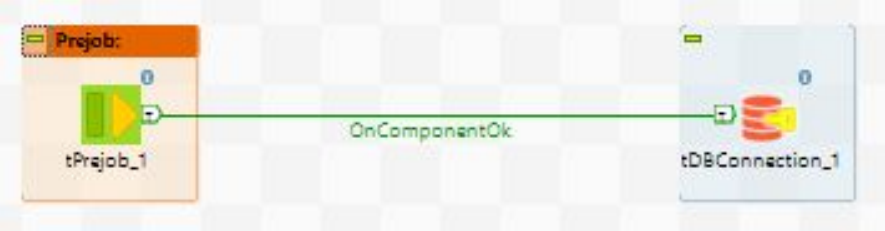
Transformação Camada Raw - Ingestão

O primeiro job, “Job ingestao” faz a conexão no banco de dados para buscar as tabelas a serem ingeridas. Tais tabelas com os dados brutos são salvas em formato csv na pasta “raw” na máquina local.





Transformação Camada Raw - Ingestão



No início do job, é feita a conexão com o banco de dados.

Component

tDBConnection_1 (Microsoft SQL Server)

- Basic settings
- Advanced settings
- Dynamic settings
- Visão
- Documentação

Database: Microsoft SQL Server — Apply
Property Type: Built-In
JDBC Provider: Open source JTDS
Host: "sqlservercentralpublic.database.windows.net" Porta: "1433" Schema: "SalesLT"
Database: "AdventureWorks"
Username: "sqlfamily" Senha: ********
Additional JDBC Parameters: ""
☐ Use or register a shared DB Connection

Data source
This option only applies when deploying and running in the Talend Runtime
☐ Specify a data source alias

Transformação Camada Raw - Ingestão

Para cada tabela, é feita a consulta dos dados via query no banco de dados, como mostra o exemplo ao lado.

Consulta_SalesOrderDetail(tDBInput_2)(Microsoft SQL Server)

Basic settings | Advanced settings | Dynamic settings | Visão | Documentação

Database: Microsoft SQL Server — Apply
Use an existing connection
Property Type: Built-In
JDBC Provider: Open source JTDS
Host: "sqlservercentralpublic.database.windows.net" Porta: "1433" Schema: "SalesLT"
Database: "AdventureWorks"
Username: "sqlfamily" Senha: ********
Schema: Repository — DB (MSSQL):BD_AdventureWorks - SalesOrd — Edit schema
Nome da Tabela: "SalesOrderDetail"
Tipo da consulta: Built-In — Guess Query — Guess schema
Query:

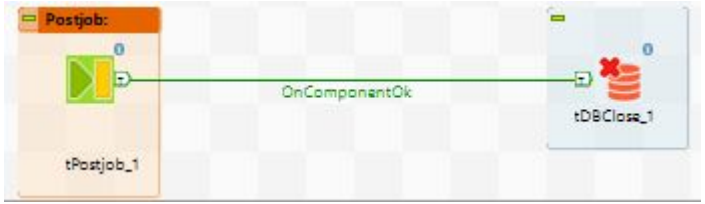
```
"SELECT SalesLT.SalesOrderDetail.SalesOrderID,
    SalesLT.SalesOrderDetail.SalesOrderDetailID,
    SalesLT.SalesOrderDetail.OrderQty,
    SalesLT.SalesOrderDetail.ProductID,
    SalesLT.SalesOrderDetail.UnitPrice,
    SalesLT.SalesOrderDetail.UnitPriceDiscount,
    SalesLT.SalesOrderDetail.LineTotal,
    SalesLT.SalesOrderDetail.rowguid,
    SalesLT.SalesOrderDetail.ModifiedDate
FROM  SalesLT.SalesOrderDetail"
```

tFileOutputDelimited_2

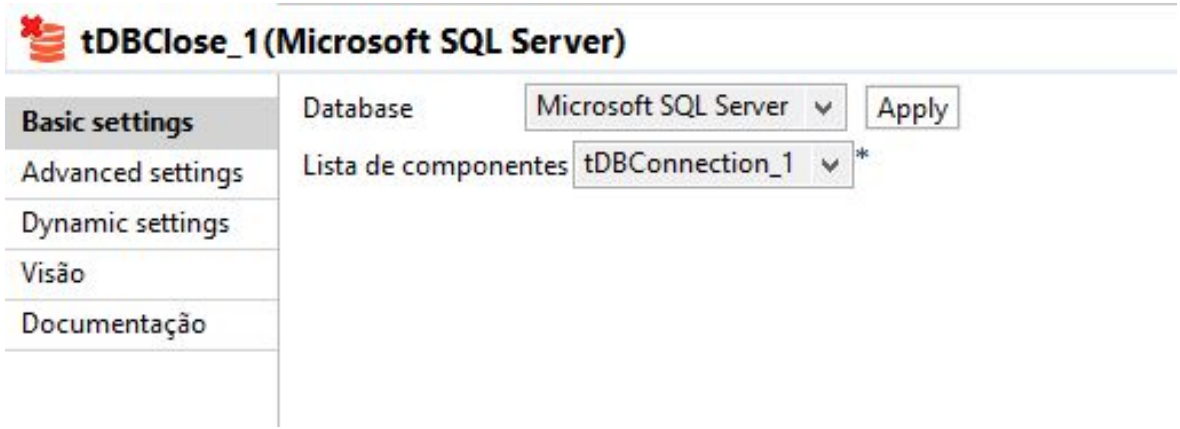
Basic settings | Advanced settings | Dynamic settings | Visão | Documentação

Property Type: Built-In
Use Output Stream
File Name: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/raw/SalesOrderDetail.csv"
Use OS line separator as row separator when CSV Row Separator is set to CR,LF or CRLF.
CSV Row Separator: LF("\n") Field Separator: ";"
Incluir no final | Incluir cabeçalho | Compress as zip file
Schema: Built-In — Edit schema — Sync columns

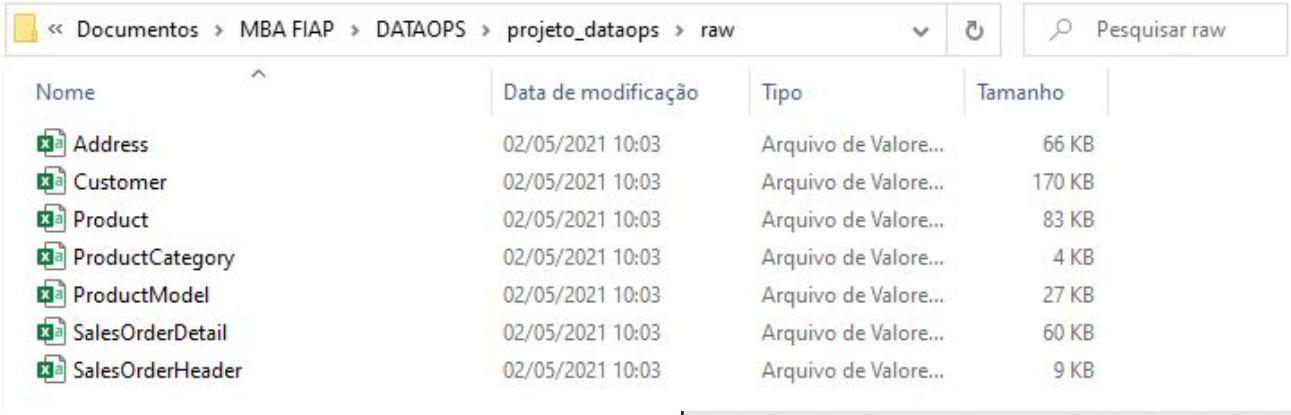
Transformação Camada Raw - Ingestão



Ao final do job, a conexão com o banco de dados é encerrada.



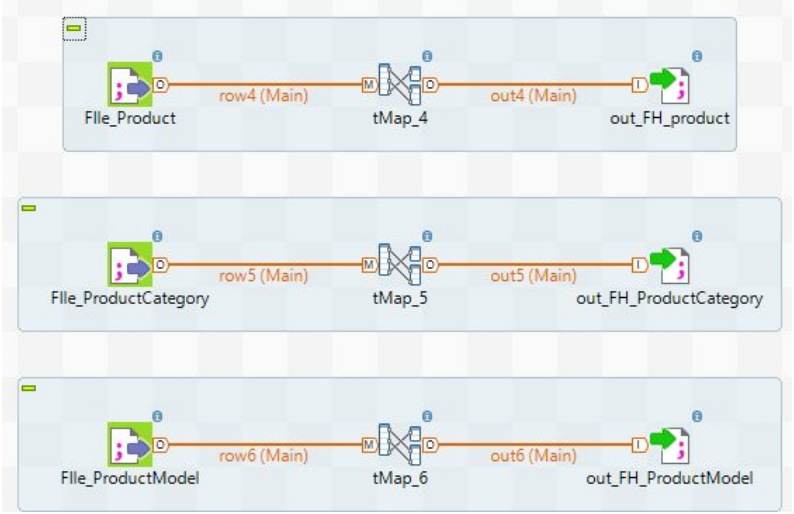
Transformação Camada Raw - Resultados



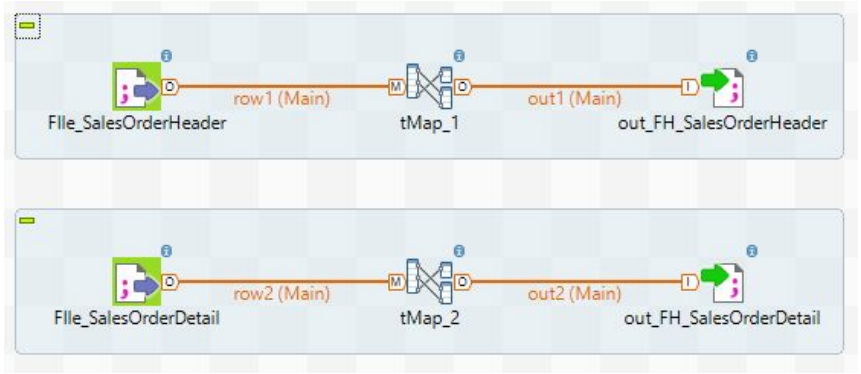
	A	B	C	D	E	F	G	H	I	J	K	L
1	ProductID	ProductNumber	Name	Color	StandardCost	ListPrice	Size	Weight	ProductCategoryID	ProductModelID	SellStartDate	SellEndDate
2	680	FR-R92B-58	HL Road Frame - Black, 58	Black	10.593.100	14.315.000	58	1016.04	18	6	Sat Jun 01 00:00:00 BRT 2002	
3	706	FR-R92R-58	HL Road Frame - Red, 58	Red	10.593.100	14.315.000	58	1016.04	18	6	Sat Jun 01 00:00:00 BRT 2002	
4	707	HL-U509-R	Sport-100 Helmet, Red	Red	130.863	349.900			35	33	Fri Jul 01 00:00:00 BRT 2005	
5	708	HL-U509	Sport-100 Helmet, Black	Black	130.863	349.900			35	33	Fri Jul 01 00:00:00 BRT 2005	
6	709	SO-B909-M	Mountain Bike Socks, M	White	33.963	95.000	M		27	18	Fri Jul 01 00:00:00 BRT 2005	Fri Jun 30 00:00:00
7	710	SO-B909-L	Mountain Bike Socks, L	White	33.963	95.000	L		27	18	Fri Jul 01 00:00:00 BRT 2005	Fri Jun 30 00:00:00
8	711	HL-U509-B	Sport-100 Helmet, Blue	Blue	130.863	349.900			35	33	Fri Jul 01 00:00:00 BRT 2005	
9	712	CA-1098	AWC Logo Cap	Multi	69.223	89.900			23	2	Fri Jul 01 00:00:00 BRT 2005	
10	713	LJ-0192-S	Long-Sleeve Logo Jersey, S	Multi	384.923	499.900	S		25	11	Fri Jul 01 00:00:00 BRT 2005	
11	714	LJ-0192-M	Long-Sleeve Logo Jersey, M	Multi	384.923	499.900	M		25	11	Fri Jul 01 00:00:00 BRT 2005	
12	715	LJ-0192-L	Long-Sleeve Logo Jersey, L	Multi	384.923	499.900	L		25	11	Fri Jul 01 00:00:00 BRT 2005	
13	716	LJ-0192-X	Long-Sleeve Logo Jersey, XL	Multi	384.923	499.900	XL		25	11	Fri Jul 01 00:00:00 BRT 2005	
14	717	FR-R92R-62	HL Road Frame - Red, 62	Red	8.686.342	14.315.000	62	1043.26	18	6	Fri Jul 01 00:00:00 BRT 2005	
15	718	FR-R92R-44	HL Road Frame - Red, 44	Red	8.686.342	14.315.000	44	961.61	18	6	Fri Jul 01 00:00:00 BRT 2005	
16	719	FR-R92R-48	HL Road Frame - Red, 48	Red	8.686.342	14.315.000	48	979.75	18	6	Fri Jul 01 00:00:00 BRT 2005	
17	720	FR-R92R-52	HL Road Frame - Red, 52	Red	8.686.342	14.315.000	52	997.90	18	6	Fri Jul 01 00:00:00 BRT 2005	
18	721	FR-R92R-56	HL Road Frame - Red, 56	Red	8.686.342	14.315.000	56	1016.04	18	6	Fri Jul 01 00:00:00 BRT 2005	
19	722	FR-R38B-58	LL Road Frame - Black, 58	Black	2.046.251	3.372.200	58	1115.83	18	9	Fri Jul 01 00:00:00 BRT 2005	
20	723	FR-R38B-60	LL Road Frame - Black, 60	Black	2.046.251	3.372.200	60	1124.90	18	9	Fri Jul 01 00:00:00 BRT 2005	
21	724	FR-R38B-62	LL Road Frame - Black, 62	Black	2.046.251	3.372.200	62	1133.98	18	9	Fri Jul 01 00:00:00 BRT 2005	
22	725	FR-R38R-44	LL Road Frame - Red, 44	Red	1.871.571	3.372.200	44	1052.33	18	9	Fri Jul 01 00:00:00 BRT 2005	Sat Jun 30 00:00:00
23	726	FR-R38R-48	LL Road Frame - Red, 48	Red	1.871.571	3.372.200	48	1070.47	18	9	Fri Jul 01 00:00:00 BRT 2005	Sat Jun 30 00:00:00
24	727	FR-R38R-52	LL Road Frame - Red, 52	Red	1.871.571	3.372.200	52	1088.62	18	9	Fri Jul 01 00:00:00 BRT 2005	Sat Jun 30 00:00:00

Transformações Camada Harmonized

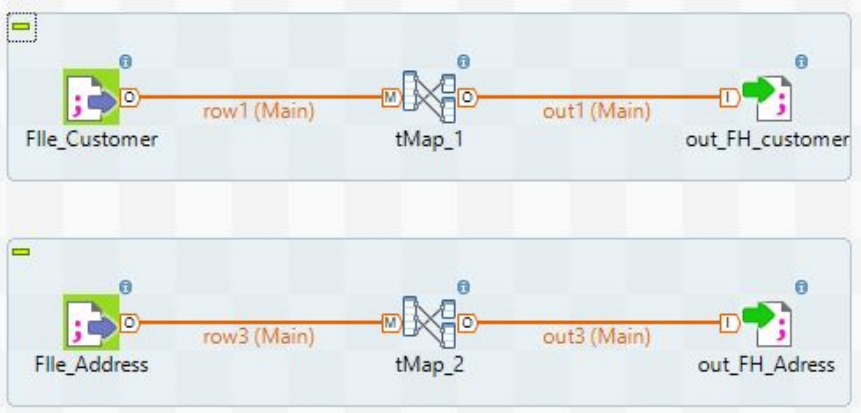
Após feita a ingestão dos dados na camada “raw”, os jobs da camada harmonized podem ser executados.
Nesta camada, dividiu-se os dados em três entidades: cliente, produto e vendas.
Aqui, os dados são formatados e é feita a seleção das colunas que irão compor os dados dessa camada.



Entidade Produto

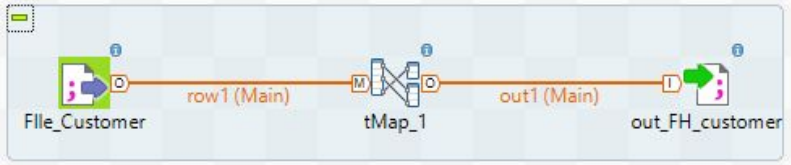


Entidade Vendas

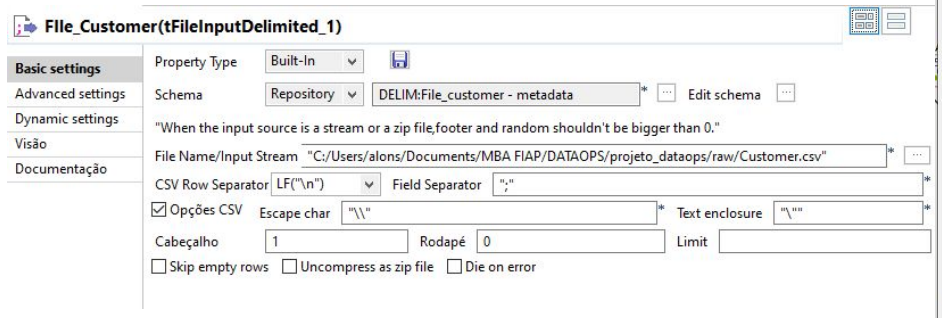


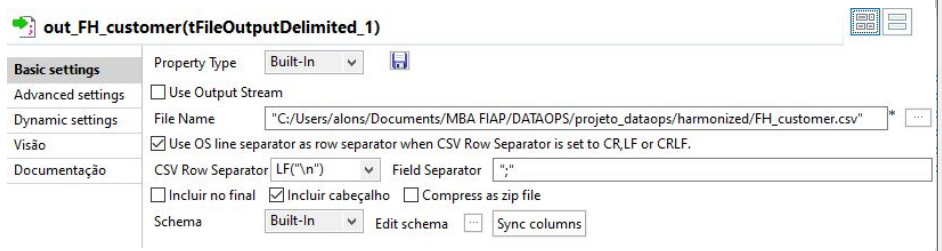
Entidade Cliente

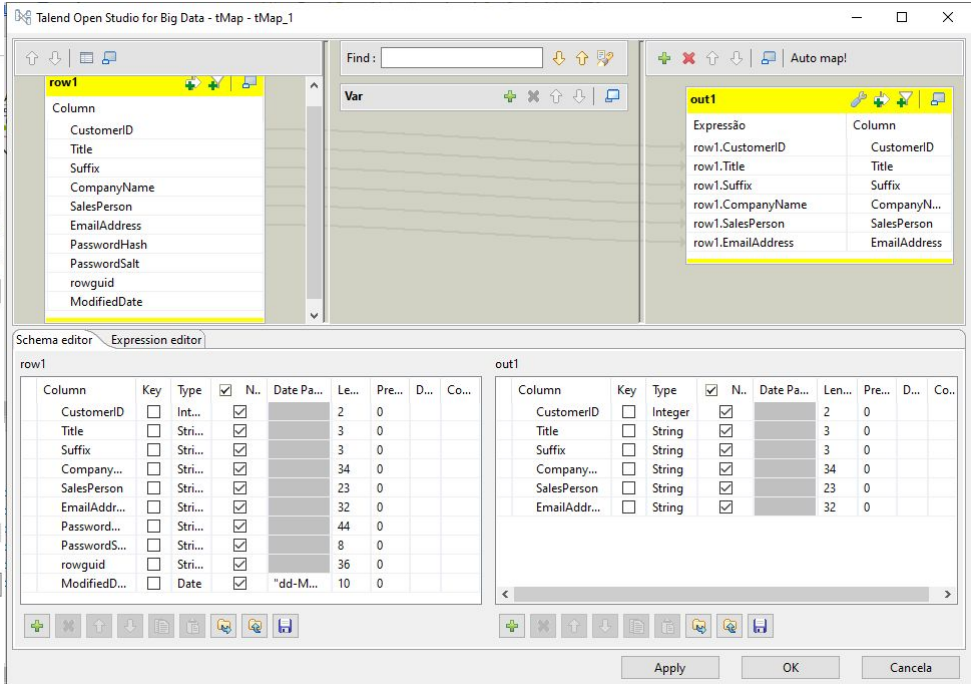
Transformação Camada Harmonized



Para cada tabela da camada “raw”, são feitas transformações e o resultado disso é salvo na camada “harmonized”. A seguir é mostrado o exemplo do que é feito em uma das tabelas.







Transformação Camada Harmonized - Resultados

	A	B	C	D	E	F	G	H	I
1	SalesOrderID	RevisionNumber	OrderDate	DueDate	ShipDate	Status	OnlineOrderFlag	SalesOrderNumber	PurchaseOrderNumber
2	71774	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71774	PO348186287
3	71776	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71776	PO19952192051
4	71780	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71780	PO19604173239
5	71782	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71782	PO19372114749
6	71783	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71783	PO19343113609
7	71784	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71784	PO19285135919
8	71796	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71796	PO17052159664
9	71797	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71797	PO16501134889
10	71815	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71815	PO13021155785
11	71816	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71816	PO12992180445
12	71831	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71831	PO10295111084
13	71832	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71832	PO10353140756
14	71845	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71845	PO2697119362
15	71846	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71846	PO2378131604
16	71856	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71856	PO16530177647
17	71858	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71858	PO16153112278
18	71863	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71863	PO16124166561
19	71867	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71867	PO13050111529
20	71885	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71885	PO6119130779
21	71895	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71895	PO3770176273
22	71897	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71897	PO4785152479
23	71898	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71898	PO5713190501
24	71899	2	01/06/2008	13/06/2008	08/06/2008	5	false	SO71899	PO4582142611

Transformação Camada Curated

Na camada “curated” são armazenados os resultados de consolidações, datasets, relatórios, entre outros produtos de dados analíticos. Para essa atividade, foram feitos alguns produtos, que vamos descrever por contexto e fase.

1. **Relatório para área de vendas.**

Envio diário. Valores de venda por região e venda online e offline.

2. **Dataset para exploração - ciência de dados.**

Envio pontual. Objetivo de construir modelo de previsão de churn. Amostra de dados com informações das entidades de cliente, produtos, vendas e vendedores.

3. **Alteração no Relatório 1.**

Envio diário. Incluir informações sobre vendedores com melhor e pior desempenho.

4. **Relatório de produtos vendidos.**

Envio mensal. Distribuição de produtos por região com métricas de quantidade de venda e valor.

5. **Dataset de produtos vendidos.**

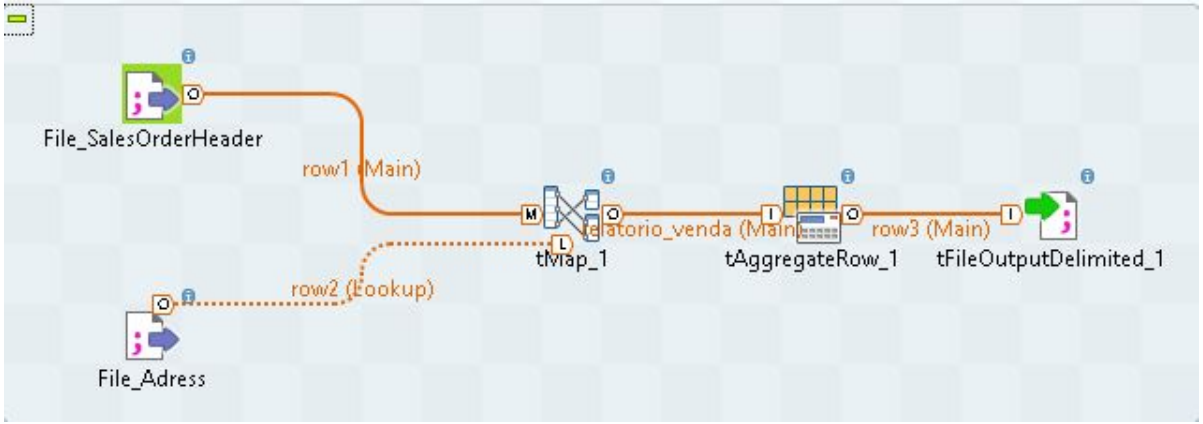
Envio semanal. Informações de produtos vendidos, vendedor que realizou, distribuído por região e data, para que o time de possa realizar análises manuais.

6. **Alteração no Dataset 2.**

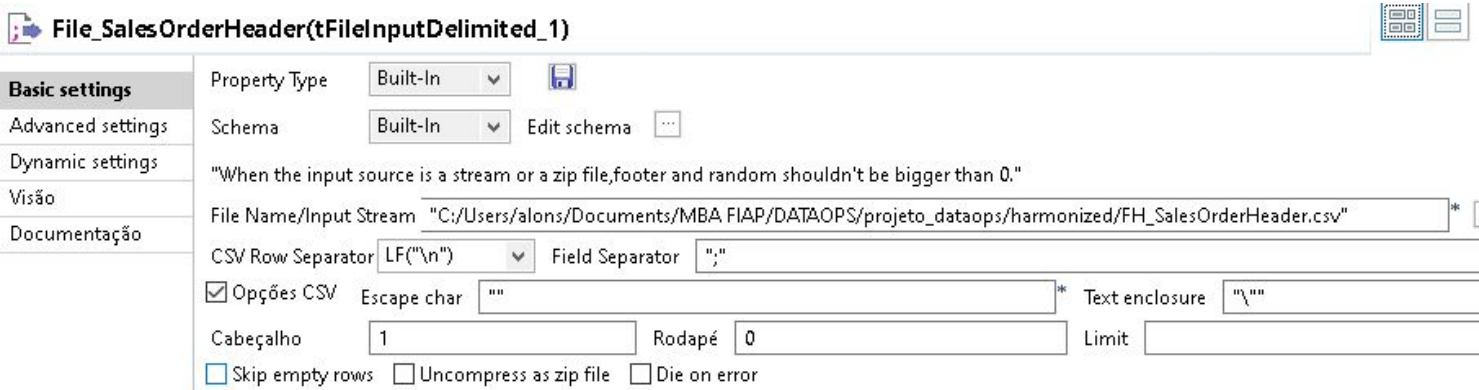
Envio pontual. Nova amostra de dados com informações das entidades de produtos, vendas e vendedores, removendo a entidade de cliente e incluindo as informações existentes sobre as lojas das vendas. A amostra deve ser enviada com um histórico maior, já dividida em amostra para modelagem e amostra para treino.

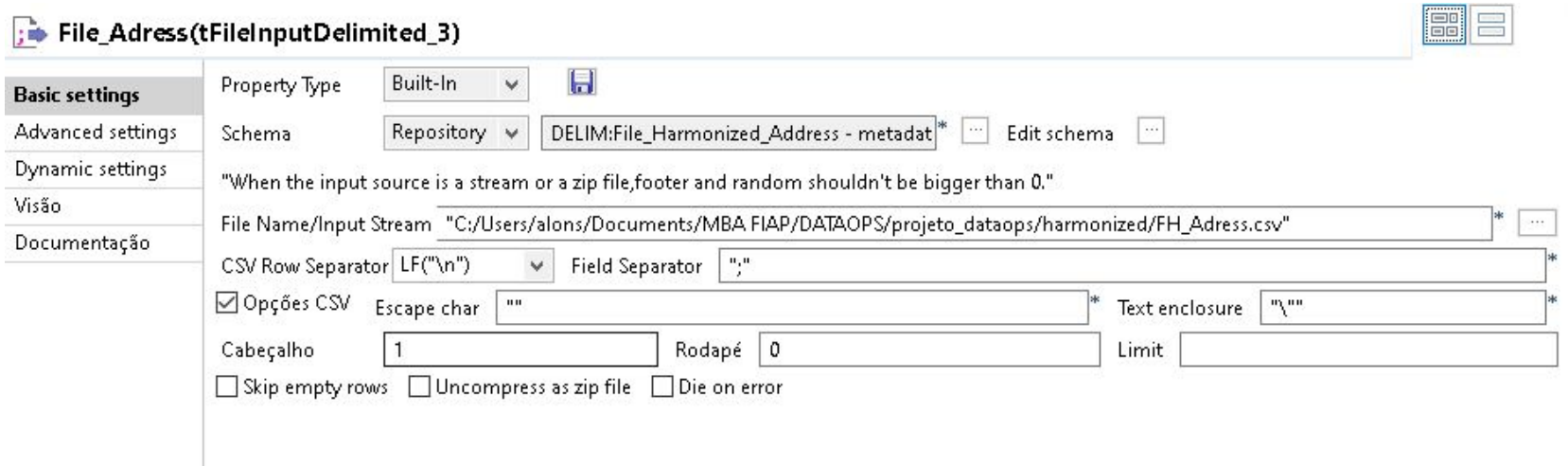
1 - Relatório para área de vendas

Nesse job, são coletados dados das tabelas SalesOrderHeader e Address.
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.
No step tAggregateRow, são feitas as agregações dos dados pelos campos de data, região e flag de venda online.
É feita também a soma do valor vendido considerando essas agregações.
No último step, o arquivo final é salvo na camada “curated”.
Nos slides a seguir são mostrados detalhes de cada step.

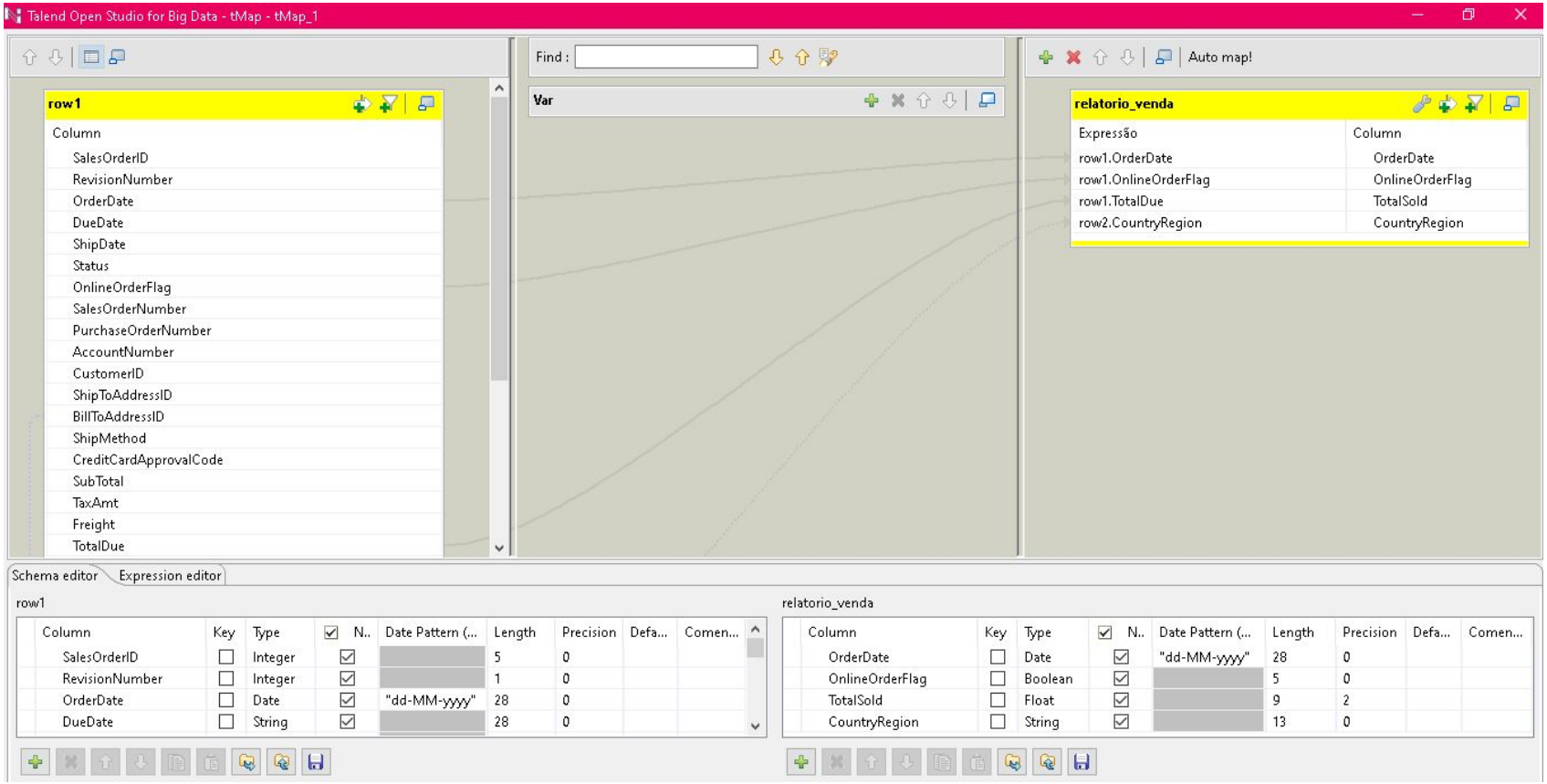


1 - Relatório para área de vendas

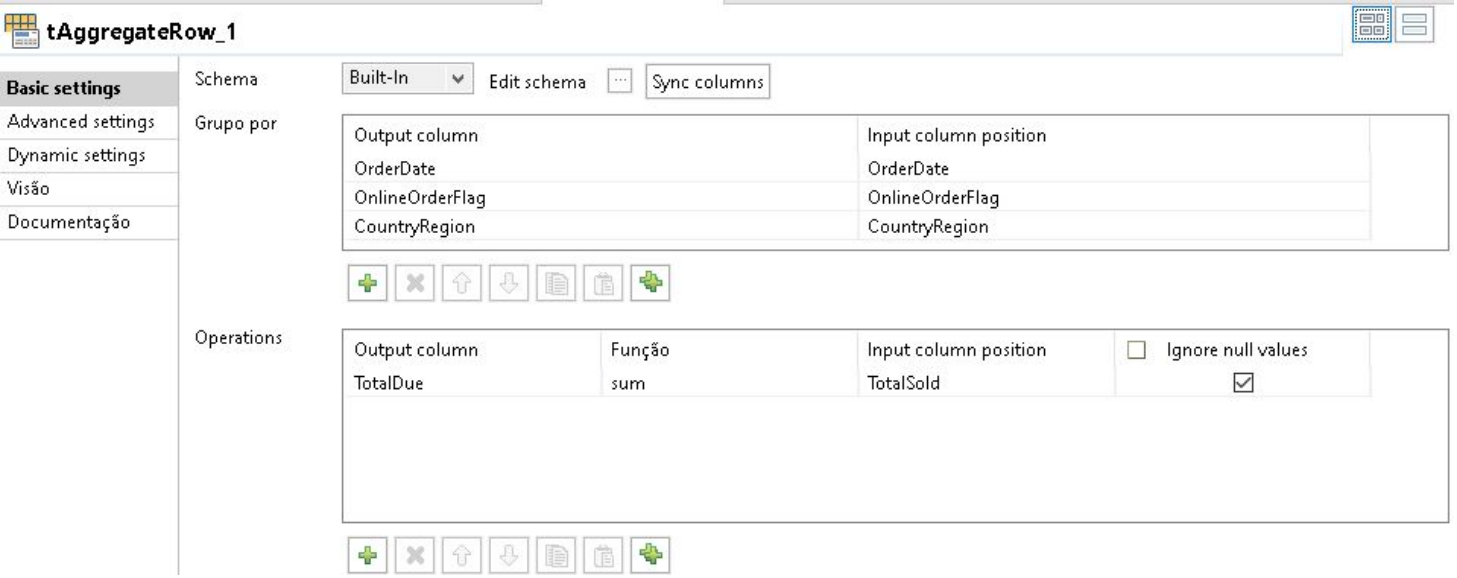


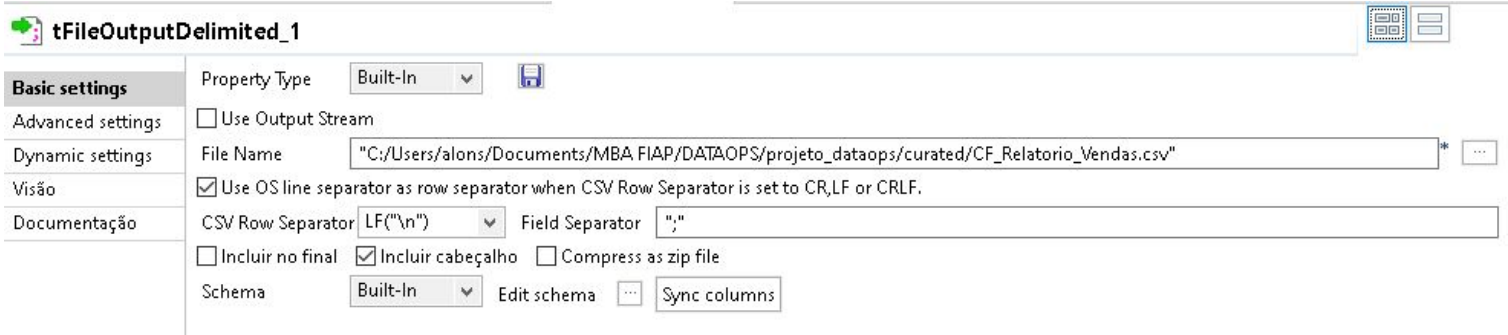


1 - Relatório para área de vendas



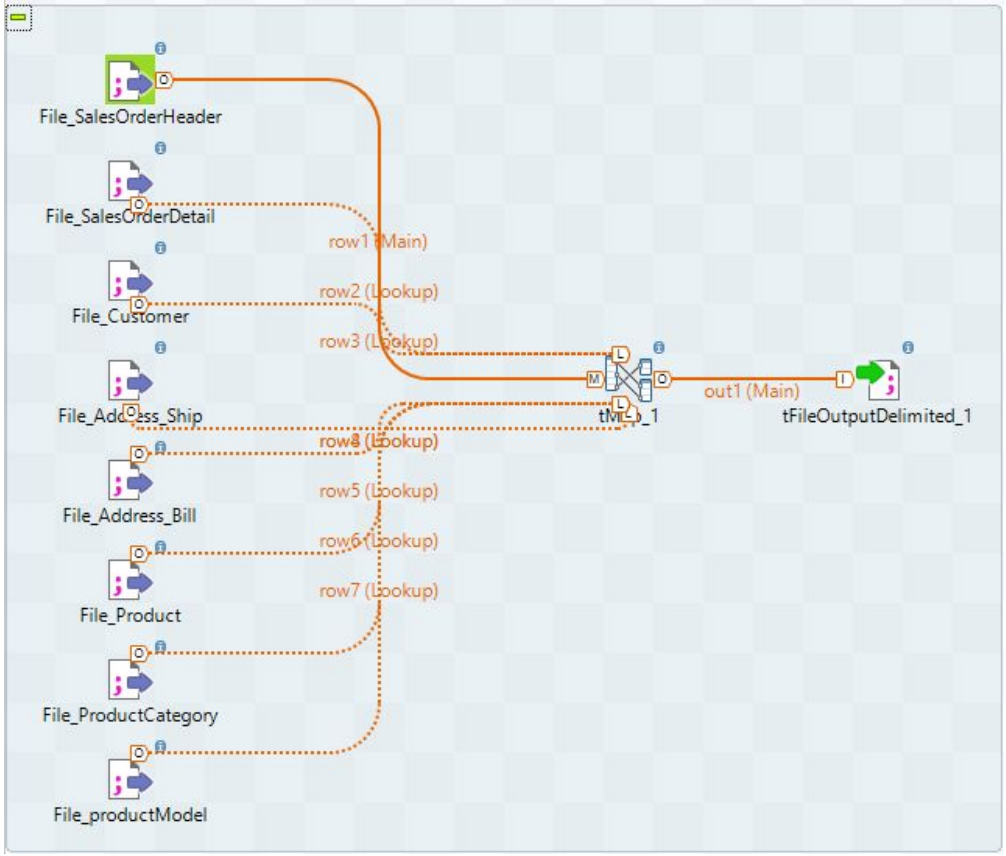
1 - Relatório para área de vendas



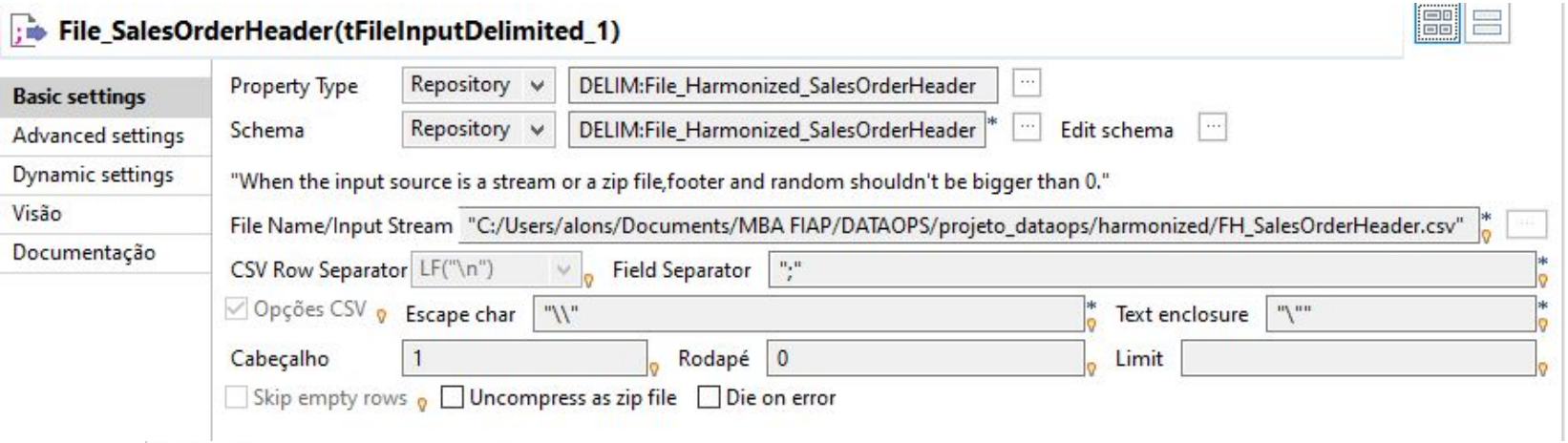


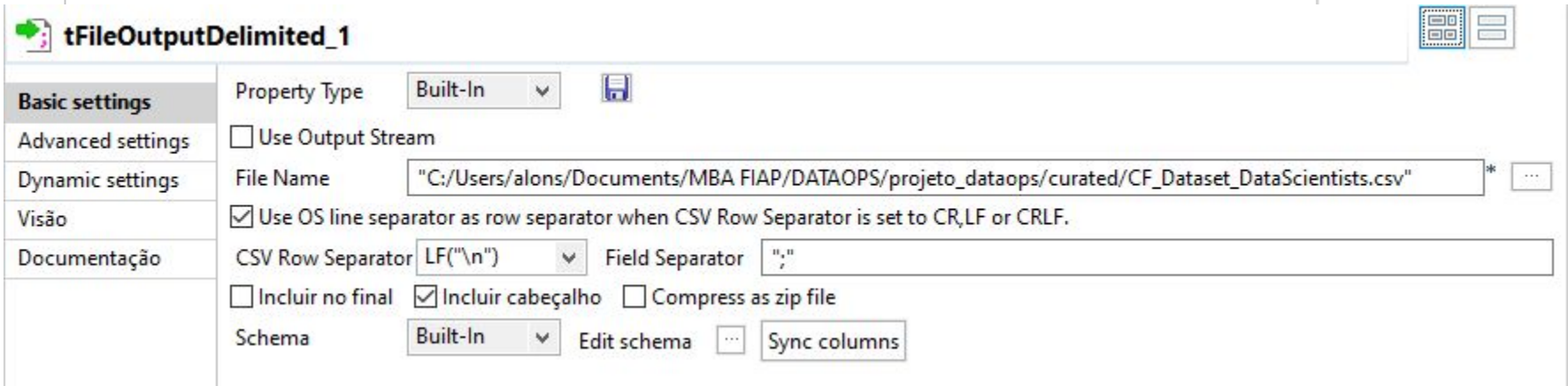
2 - Dataset para exploração - ciência de dados

Nesse job, são coletados dados das tabelas das entidades de vendas, vendedores, cliente e produto, conforme solicitado.
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.
No último step, o arquivo final é salvo na camada “curated”.
Nos slides a seguir são mostrados detalhes de cada step.

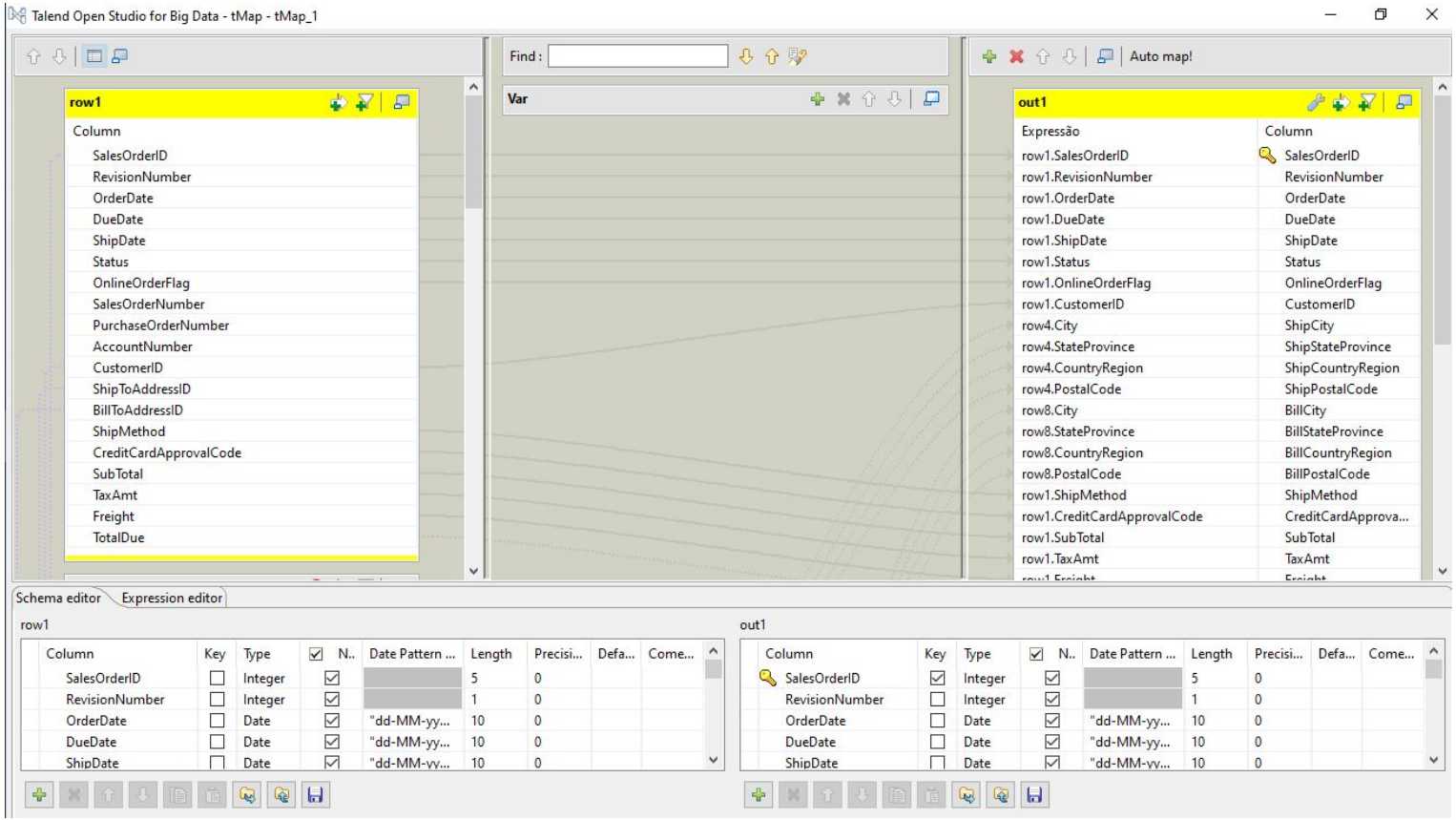


2 - Dataset para exploração - ciência de dados





2 - Dataset para exploração - ciência de dados

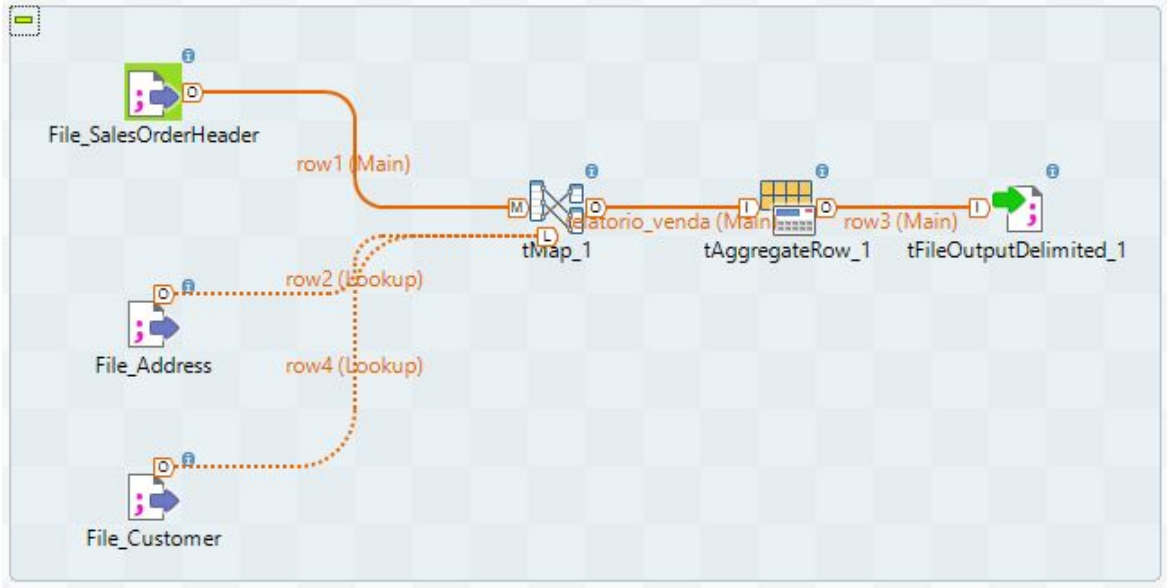


2 - Dataset para exploração - ciência de dados - Resultado

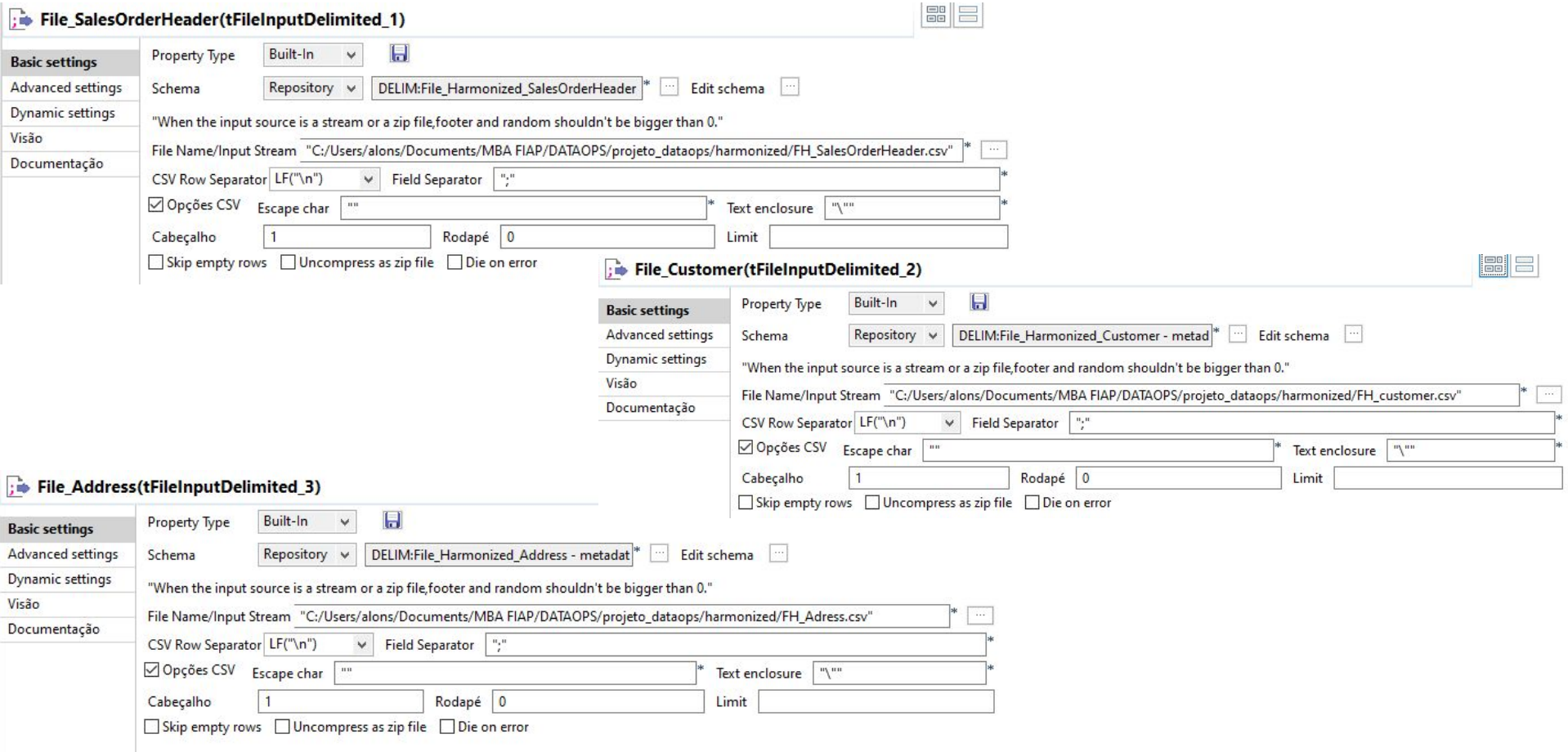
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	SalesOrderID	RevisionNumber	OrderDate	DueDate	ShipDate	Status	OnlineOrderFlag	CustomerID	ShipCity	ShipStateProvince	ShipCountryRegion	ShipPostalCode	BillCity	BillS
2	71774	2	01/06/2008	13/06/2008	08/06/2008	5	false	29847	Auburn	California	United States	95603	Auburn	Calif
3	71776	2	01/06/2008	13/06/2008	08/06/2008	5	false	30072	Wokingham	England	United Kingdom	RG41 1QW	Wokingham	Engl
4	71780	2	01/06/2008	13/06/2008	08/06/2008	5	false	30113	West Sussex	England	United Kingdom	RH15 9UD	West Sussex	Engl
5	71782	2	01/06/2008	13/06/2008	08/06/2008	5	false	29485	Van Nuys	California	United States	91411	Van Nuys	Calif
6	71783	2	01/06/2008	13/06/2008	08/06/2008	5	false	29957	Union City	California	United States	94587	Union City	Calif
7	71784	2	01/06/2008	13/06/2008	08/06/2008	5	false	29736	Woolston	England	United Kingdom	WA1 4SY	Woolston	Engl
8	71796	2	01/06/2008	13/06/2008	08/06/2008	5	false	29660	Sherman Oaks	California	United States	91403	Sherman Oaks	Calif
9	71797	2	01/06/2008	13/06/2008	08/06/2008	5	false	29796	Liverpool	England	United Kingdom	L4 4HB	Liverpool	Engl
10	71815	2	01/06/2008	13/06/2008	08/06/2008	5	false	30089	Oxnard	California	United States	93030	Oxnard	Calif
11	71816	2	01/06/2008	13/06/2008	08/06/2008	5	false	30027	Oxnard	California	United States	93030	Oxnard	Calif
12	71831	2	01/06/2008	13/06/2008	08/06/2008	5	false	30019	Milton Keynes	England	United Kingdom	MK8 8DF	Milton Keynes	Engl
13	71832	2	01/06/2008	13/06/2008	08/06/2008	5	false	29922	Milton Keynes	England	United Kingdom	MK8 8ZD	Milton Keynes	Engl
14	71845	2	01/06/2008	13/06/2008	08/06/2008	5	false	29938	Cerritos	California	United States	90703	Cerritos	Calif
15	71846	2	01/06/2008	13/06/2008	08/06/2008	5	false	30102	Cambridge	England	United Kingdom	CB4 4BZ	Cambridge	Engl
16	71856	2	01/06/2008	13/06/2008	08/06/2008	5	false	30033	Sandy	Utah	United States	84070	Sandy	Utah
17	71858	2	01/06/2008	13/06/2008	08/06/2008	5	false	29653	Santa Fe	New Mexico	United States	87501	Santa Fe	New
18	71863	2	01/06/2008	13/06/2008	08/06/2008	5	false	29975	Santa Ana	California	United States	92701	Santa Ana	Calif
19	71867	2	01/06/2008	13/06/2008	08/06/2008	5	false	29644	Oxon	England	United Kingdom	OX16 8RS	Oxon	Engl
20	71885	2	01/06/2008	13/06/2008	08/06/2008	5	false	29612	High Wycombe	England	United Kingdom	HP10 9QY	High Wycombe	Engl
21	71895	2	01/06/2008	13/06/2008	08/06/2008	5	false	29584	Culver City	California	United States	90232	Culver City	Calif
22	71897	2	01/06/2008	13/06/2008	08/06/2008	5	false	29877	Englewood	Colorado	United States	80110	Englewood	Colo
23	71898	2	01/06/2008	13/06/2008	08/06/2008	5	false	29932	Gloucestershire	England	United Kingdom	GL7 1RY	Gloucestershire	Engl
24	71899	2	01/06/2008	13/06/2008	08/06/2008	5	false	29568	El Segundo	California	United States	90245	El Segundo	Calif
25	71902	2	01/06/2008	13/06/2008	08/06/2008	5	false	29929	Fullerton	California	United States	92831	Fullerton	Calif

3 - Alteração no Relatório 1

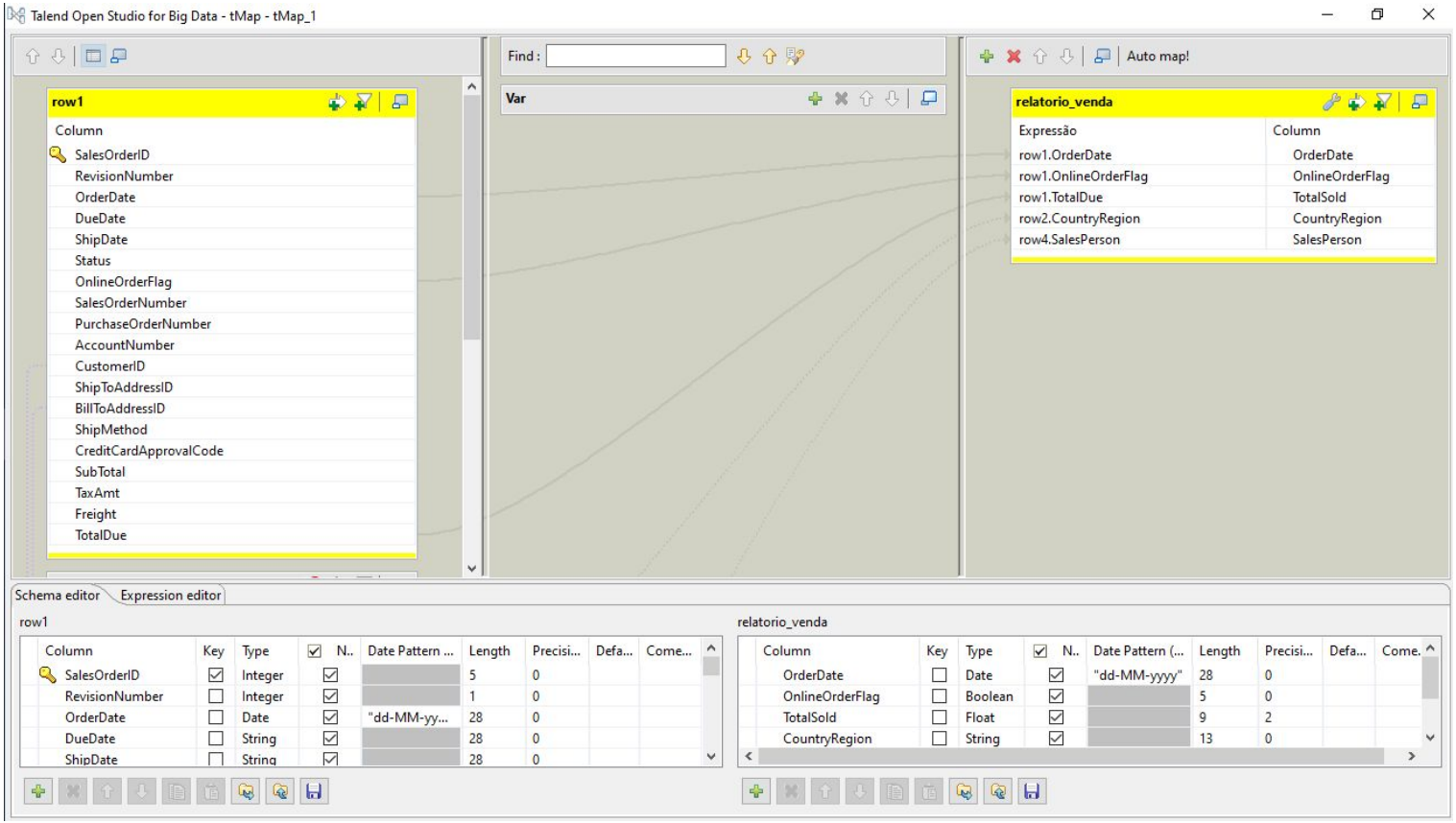
Nesse job, são coletados dados das tabelas SalesOrderHeader, Address e Customer.
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.
No step tAggregateRow, são feitas as agregações dos dados pelos campos de data, região, flag de venda online e vendedor.
É feita também a soma do valor vendido considerando essas agregações.
No último step, o arquivo final é salvo na camada “curated”.
Nos slides a seguir são mostrados detalhes de cada step.



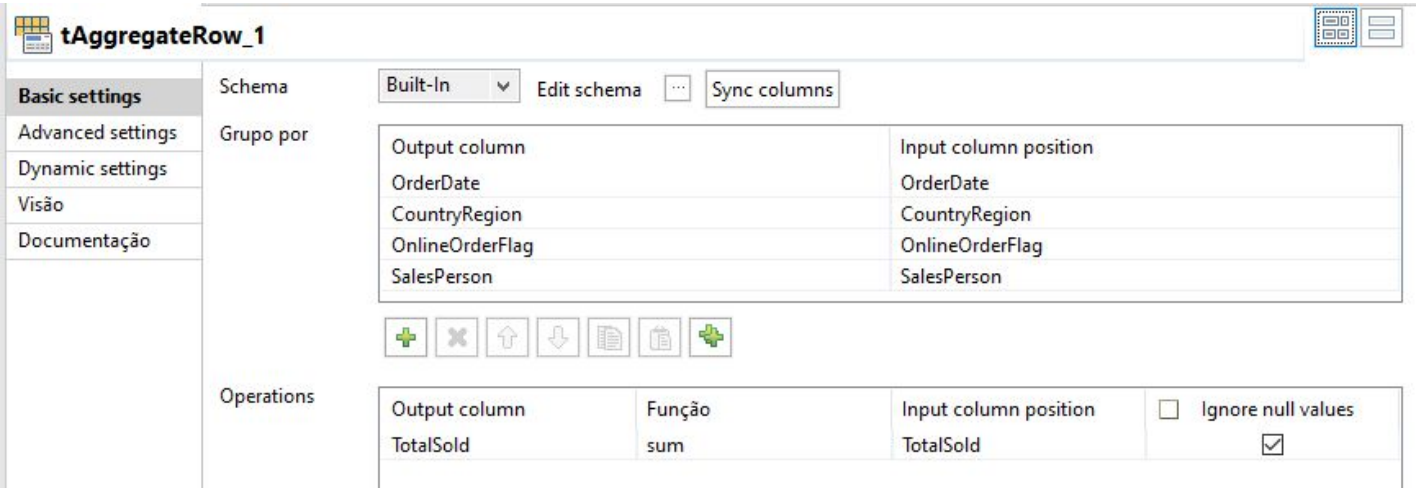
3 - Alteração no Relatório 1

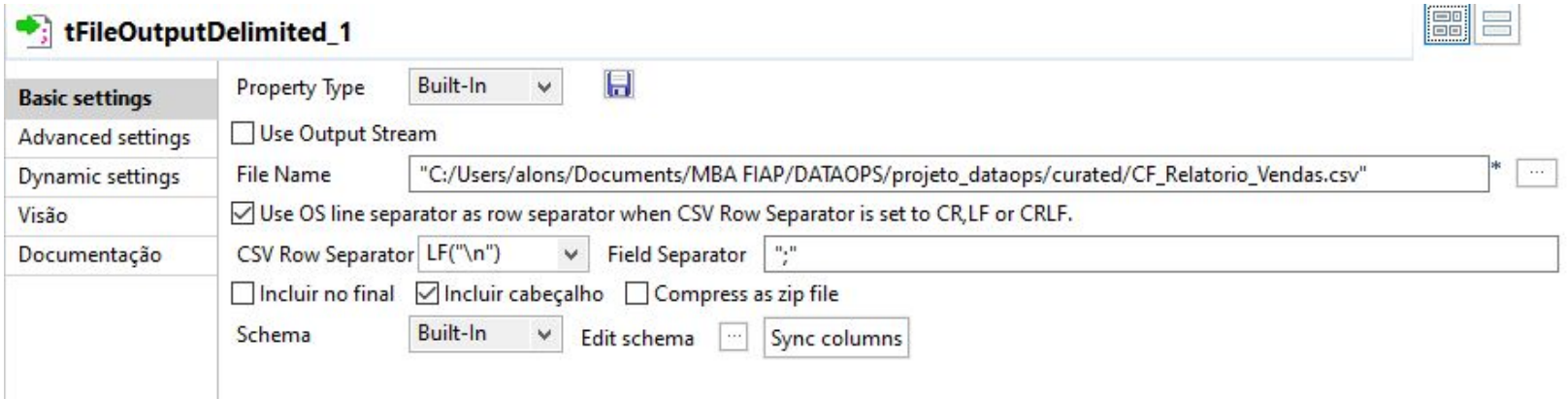


3 - Alteração no Relatório 1



3 - Alteração no Relatório 1



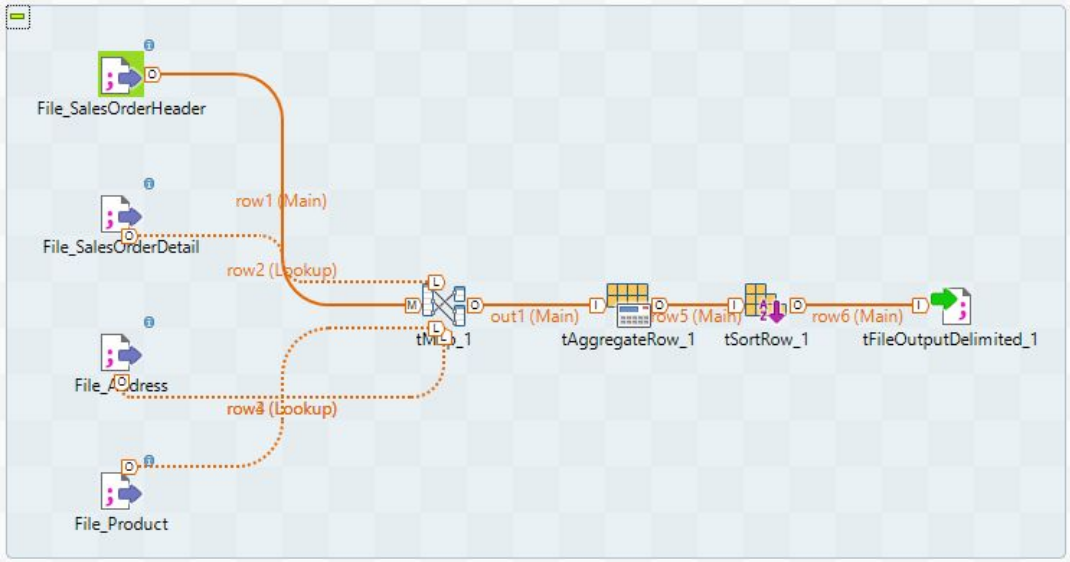


3 - Alteração no Relatório 1 - Resultado

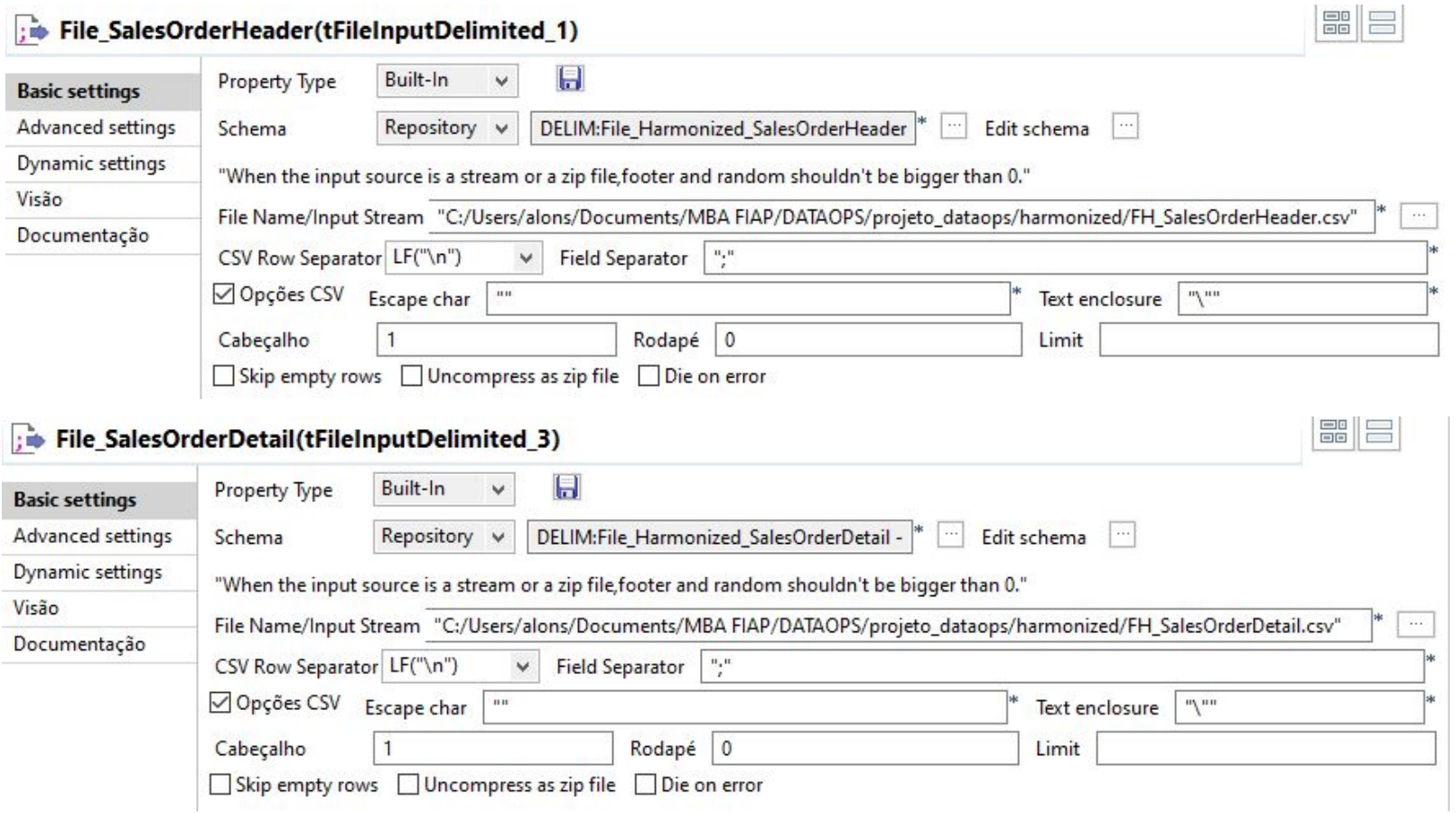
	A	B	C	D	E	F
1	OrderDate	OnlineOrderFlag	TotalSold	CountryRegion	SalesPerson	
2	01/06/2008	false	152619.12	United States	adventure-worksshu0	
3	01/06/2008	false	572496.56	United Kingdom	adventure-worksjae0	
4	01/06/2008	false	231187.9	United States	adventure-workslinda3	
5						

4 - Relatório de produtos vendidos

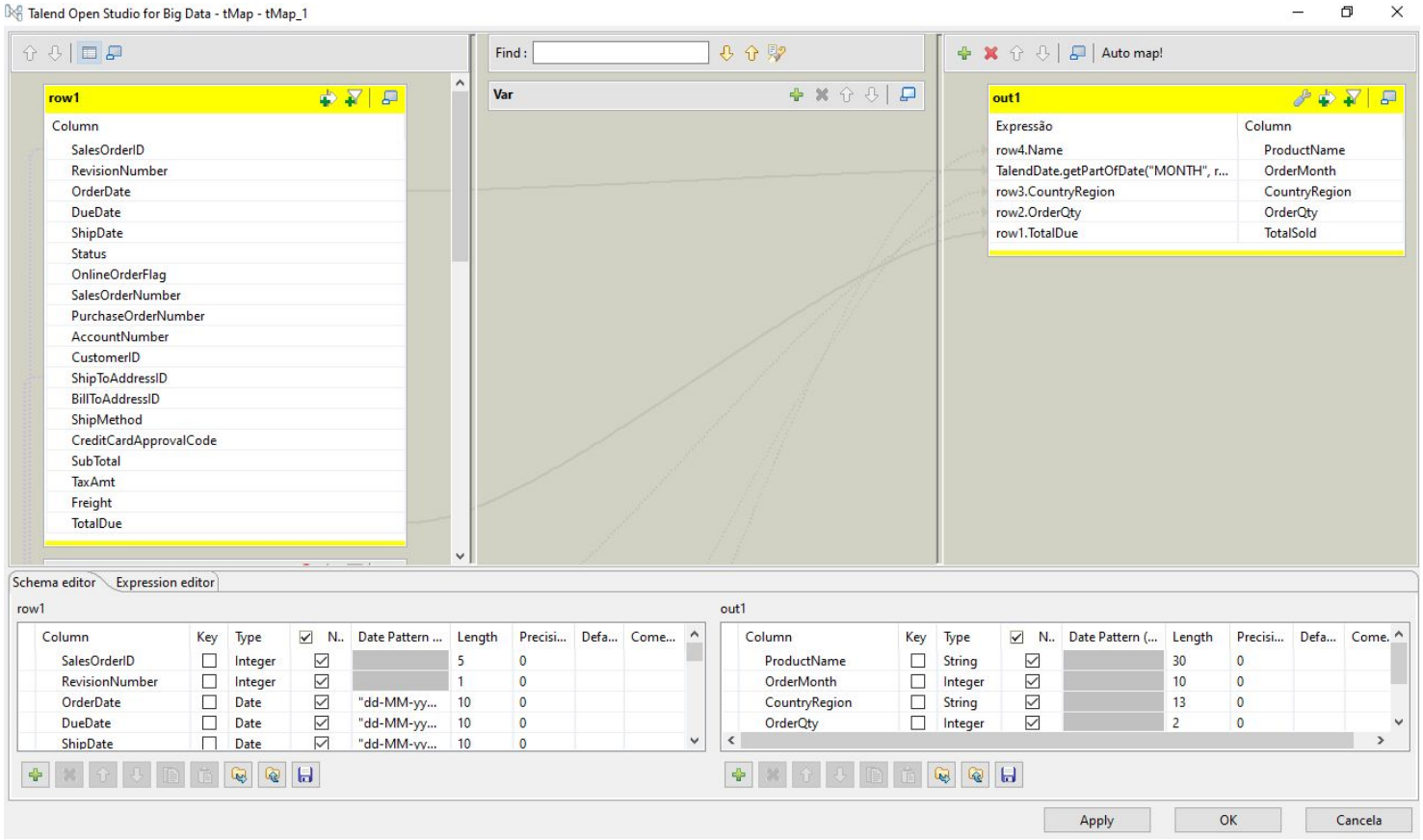
Nesse job, são coletados dados das tabelas SalesOrderHeader, SalesOrderDetail, Address e Product.
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.
No step tAggregateRow, são feitas as agregações dos dados pelos campos de nome do produto, mês de venda e região.
É feita também a soma do valor e quantidade vendidos considerando essas agregações.
Além disso, também é feito o ordenamento dos campos de nome do produto, mês de venda e quantidade vendida.
No último step, o arquivo final é salvo na camada “curated”.
Nos slides a seguir são mostrados detalhes de cada step.



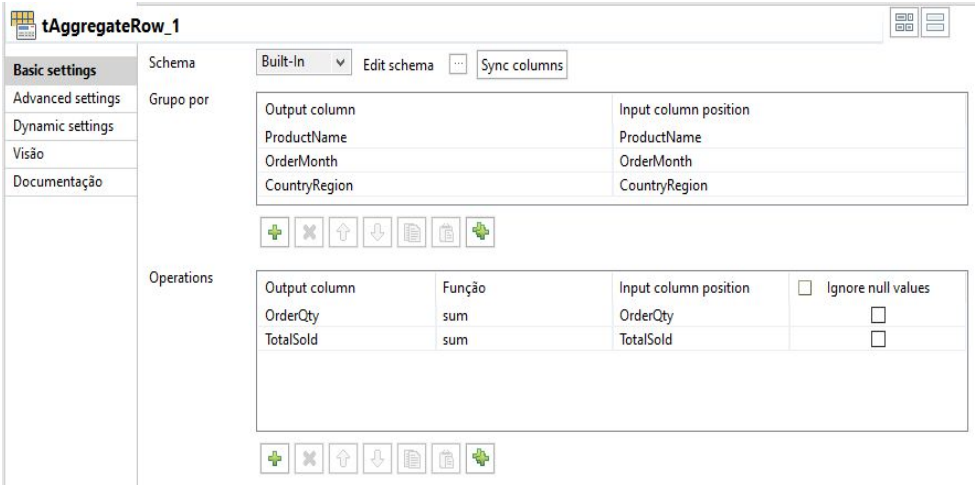
4 - Relatório de produtos vendidos

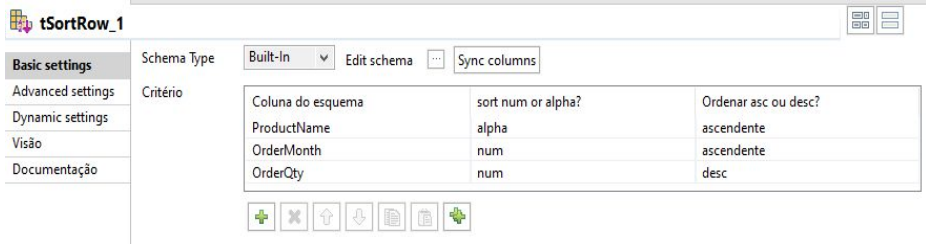


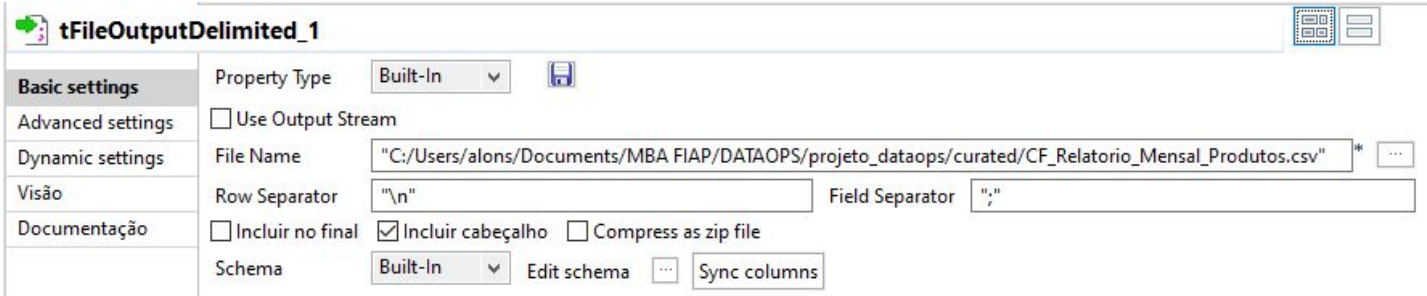
4 - Relatório de produtos vendidos



4 - Relatório de produtos vendidos



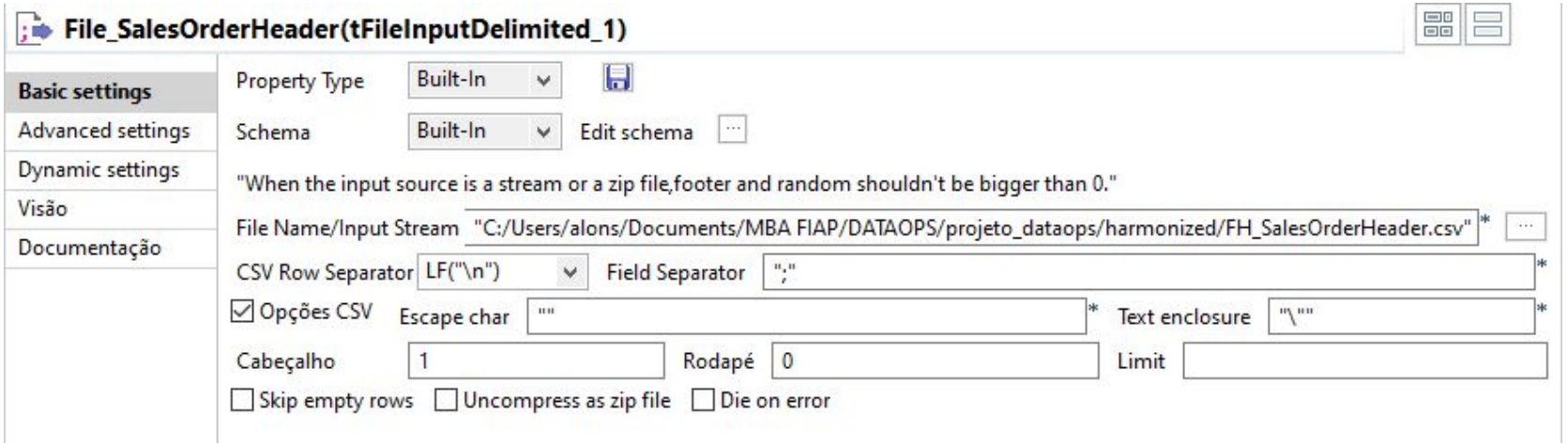


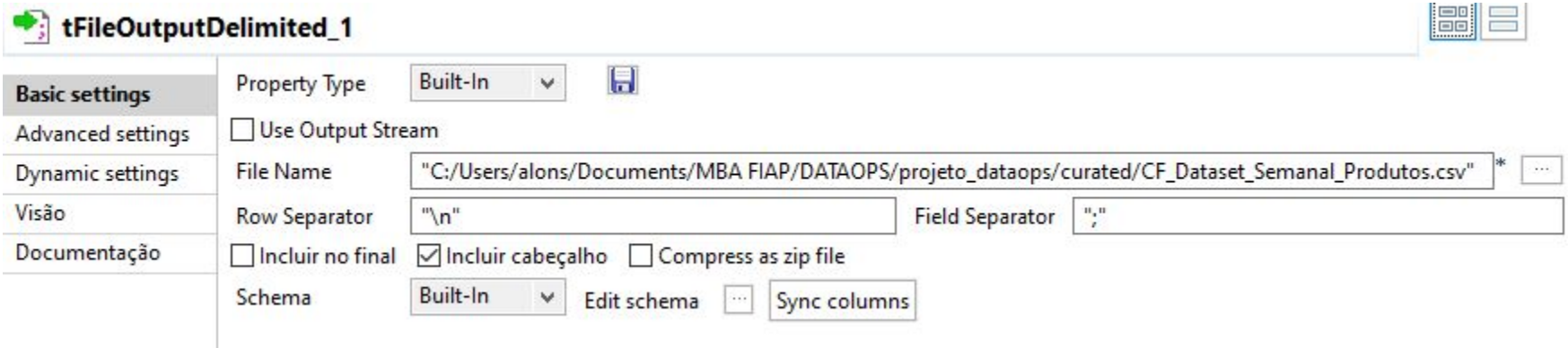


4 - Relatório de produtos vendidos - Resultado

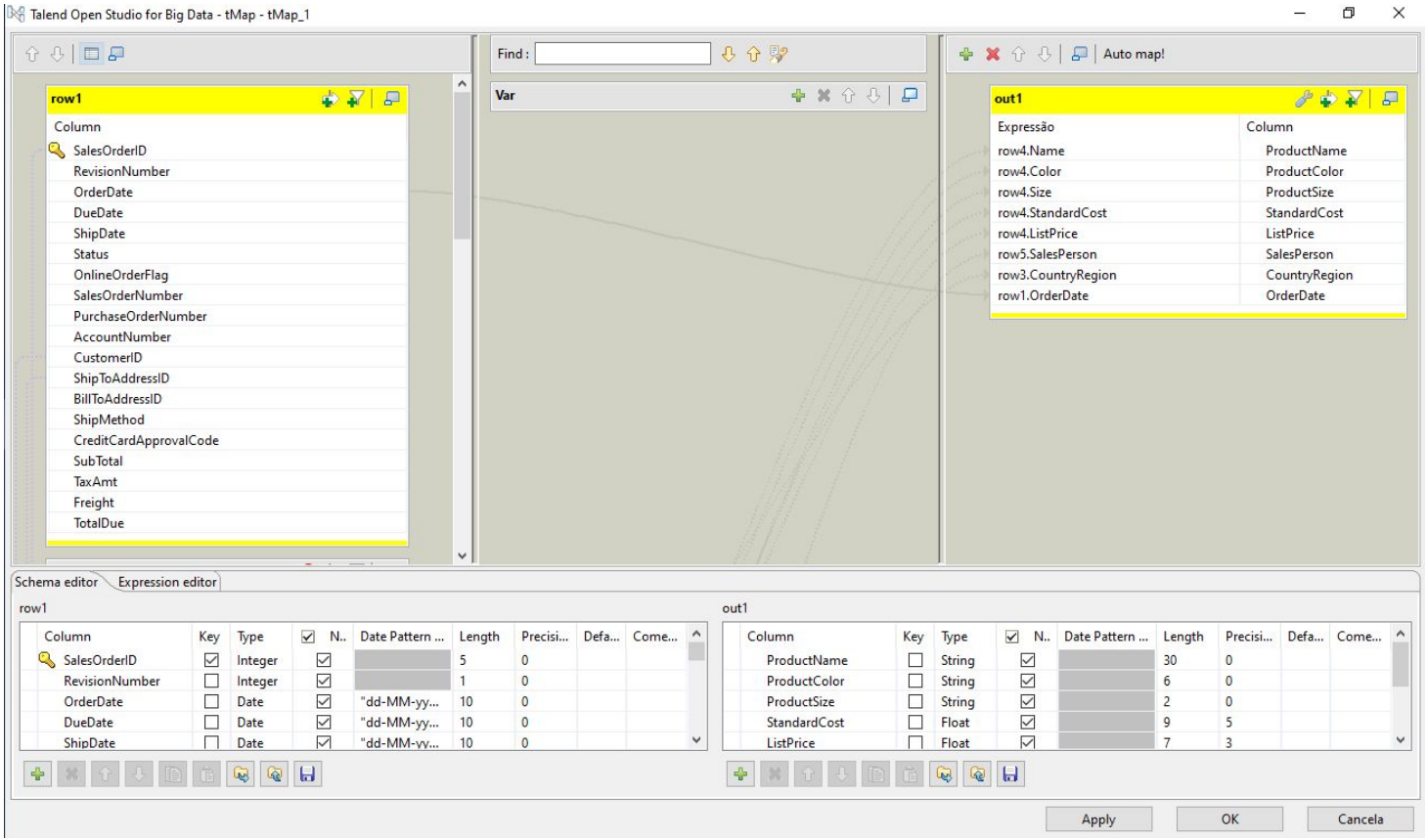
	A	B	C	D	E
1	ProductName	OrderMonth	CountryRegion	OrderQty	TotalSold
2	AWC Logo Cap	5	United States	3	15.275.197
3	Chain	5	United States	3	45.992.367
4	Classic Vest, M	5	United States	6	81.834.984
5	Classic Vest, S	5	United Kingdom	3	6.081.766
6	Classic Vest, S	5	United States	3	2.726.468
7	Front Brakes	5	United States	2	17772.88
8	Front Derailleur	5	United States	1	6.654.251
9	HL Mountain Pedal	5	United States	1	26.693.184
10	HL Road Frame - Red, 62	5	United Kingdom	1	11.705.376
11	HL Road Pedal	5	United States	1	23.616.404
12	HL Touring Frame - Yellow, 60	5	United States	2	63686.27
13	HL Touring Seat/Saddle	5	United Kingdom	1	430.437
14	Hitch Rack - 4-Bike	5	United States	1	32.937.761
15	Hydration Pack - 70 oz.	5	United Kingdom	1	42.452.652
16	LL Mountain Frame - Silver, 52	5	United Kingdom	1	39531.61
17	LL Road Pedal	5	United States	2	3.673.325
18	ML Mountain Seat/Saddle	5	United Kingdom	6	108597.95
19	ML Road Frame-W - Yellow, 38	5	United Kingdom	2	86.222.805
20	ML Road Frame-W - Yellow, 38	5	United States	1	972.785
21	Mountain-400-W Silver, 40	5	United Kingdom	2	2711.41
22	Racing Socks, L	5	United States	27	74.486.245
23	Racing Socks, L	5	United Kingdom	7	451.995
24	Racing Socks, M	5	United States	2	1.261.444

5 - Dataset de produtos vendidos





5 - Dataset de produtos vendidos



5 - Dataset de produtos vendidos - Resultado

	A	B	C	D	E	F	G	H
1	ProductName	ProductColor	ProductSize	StandardCost	ListPrice	SalesPerson	CountryRegion	OrderDate
2	ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-workslinda3	United States	01/06/2008
3	Rear Brakes	Silver		47.286	106.5	adventure-worksjae0	United Kingdom	01/06/2008
4	Hydration Pack - 70 oz.	Silver	70	205.663	54.99	adventure-worksjae0	United Kingdom	01/06/2008
5	Sport-100 Helmet, Red	Red		130.863	34.99	adventure-workslinda3	United States	01/06/2008
6	Sport-100 Helmet, Black	Black		130.863	34.99	adventure-workslinda3	United States	01/06/2008
7	Sport-100 Helmet, Black	Black		130.863	34.99	adventure-worksjae0	United Kingdom	01/06/2008
8	HL Touring Frame - Yellow, 60	Yellow	60	6.017.437	1003.91	adventure-workslinda3	United States	01/06/2008
9	ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-worksjae0	United Kingdom	01/06/2008
10	Racing Socks, M	White	M	33.623	8.99	adventure-workslinda3	United States	01/06/2008
11	Front Brakes	Silver		47.286	106.5	adventure-worksshu0	United States	01/06/2008
12	Women's Mountain Shorts, S	Black	S	261.763	69.99	adventure-worksjae0	United Kingdom	01/06/2008
13	LL Mountain Frame - Silver, 52	Silver	52	1.445.938	264.05	adventure-worksjae0	United Kingdom	01/06/2008
14	Chain	Silver		89.866	20.24	adventure-worksshu0	United States	01/06/2008
15	Mountain-400-W Silver, 40	Silver	40	4.197.784	769.49	adventure-worksjae0	United Kingdom	01/06/2008
16	Front Derailleur	Silver		406.216	91.49	adventure-worksshu0	United States	01/06/2008
17	AWC Logo Cap	Multi		69.223	8.99	adventure-workslinda3	United States	01/06/2008
18	LL Road Pedal	Silver/Black		179.776	40.49	adventure-workslinda3	United States	01/06/2008
19	HL Road Frame - Red, 62	Red	62	8.686.342	1431.5	adventure-worksjae0	United Kingdom	01/06/2008
20	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksjae0	United Kingdom	01/06/2008
21	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksshu0	United States	01/06/2008
22	Front Brakes	Silver		47.286	106.5	adventure-worksshu0	United States	01/06/2008
23	Touring-3000 Yellow, 44	Yellow	44	4.614.448	742.35	adventure-worksjae0	United Kingdom	01/06/2008
24	HL Mountain Pedal	Silver/Black		359.596	80.99	adventure-worksshu0	United States	01/06/2008

6 - Alteração no Dataset 2

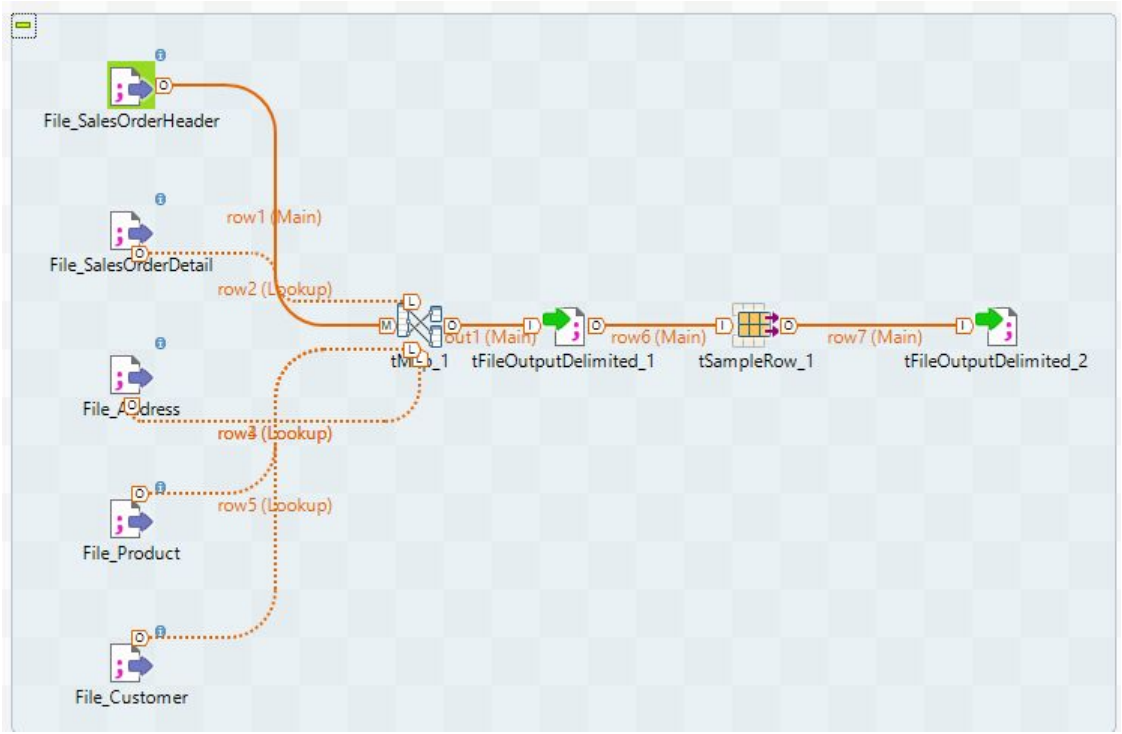
Nesse job, são coletados dados das tabelas das entidades de vendas, vendedores e produto, conforme solicitado.

No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.

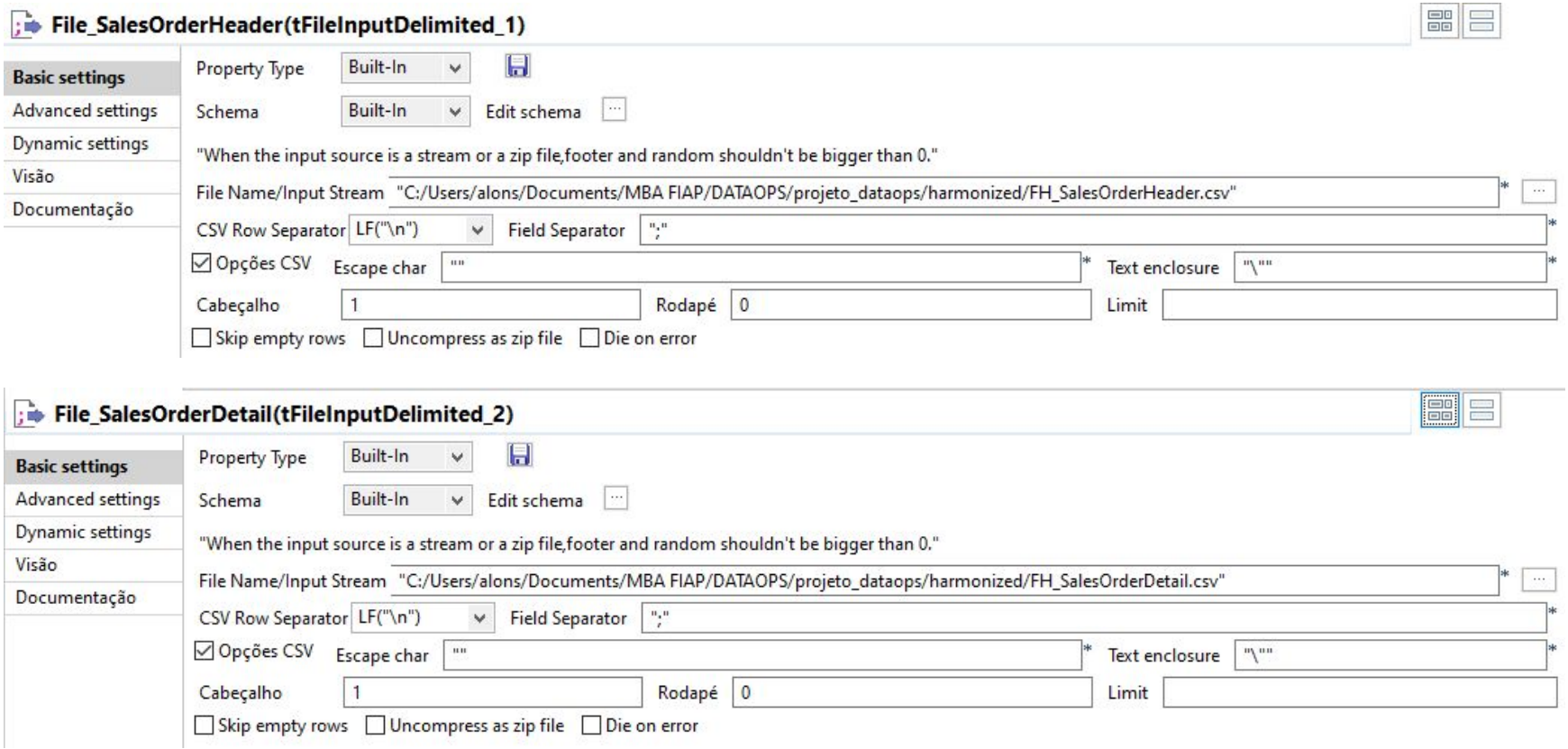
No primeiro step tFileOutputDelimited, o arquivo final com todos os registros da base (dataset para modelagem) é salvo na camada “curated”.

Em seguida, o step tSampleRow seleciona os registros para treinamento do modelo e os salva em outro arquivo, no último step.

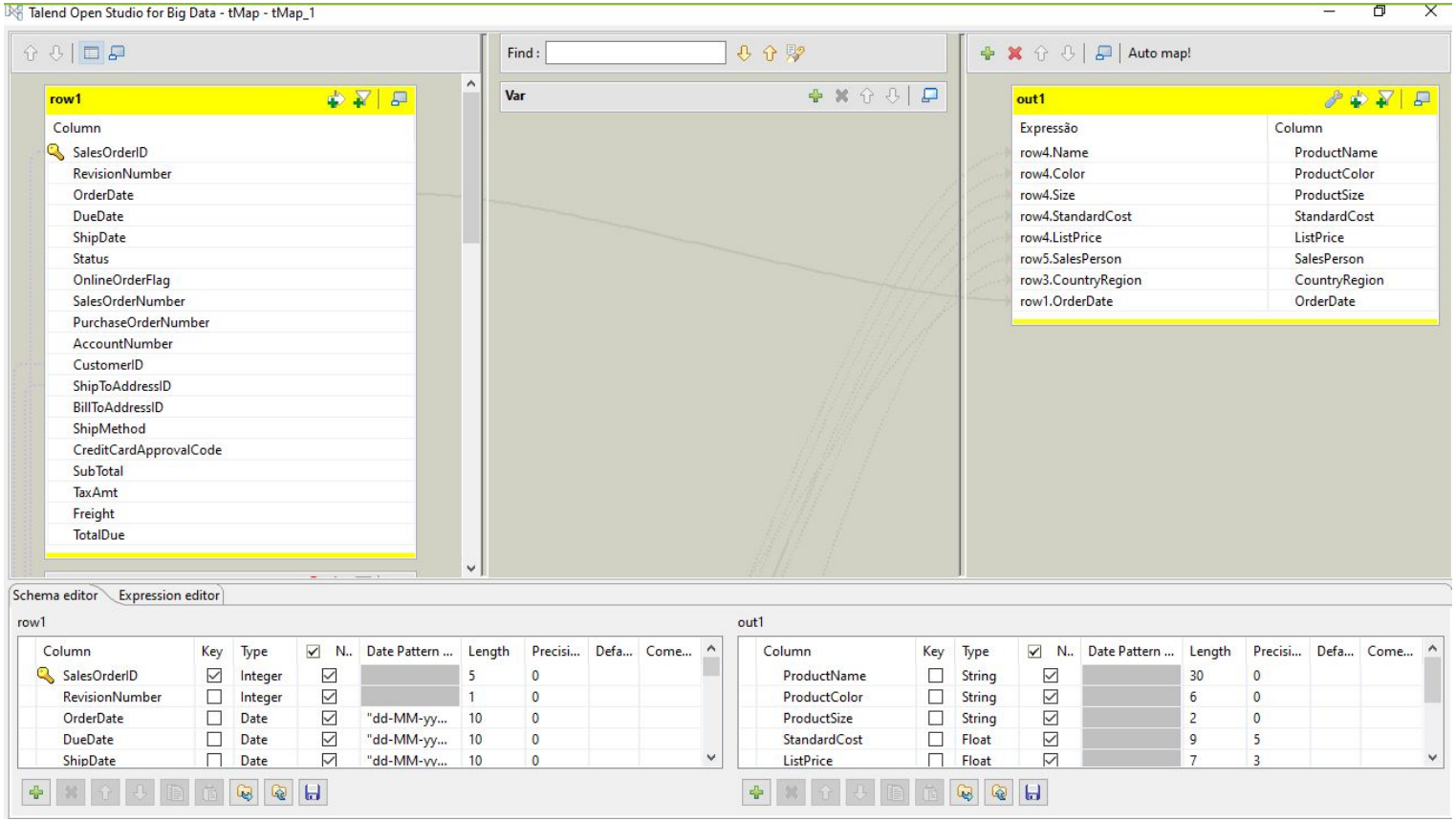
Nos slides a seguir são mostrados detalhes de cada step.



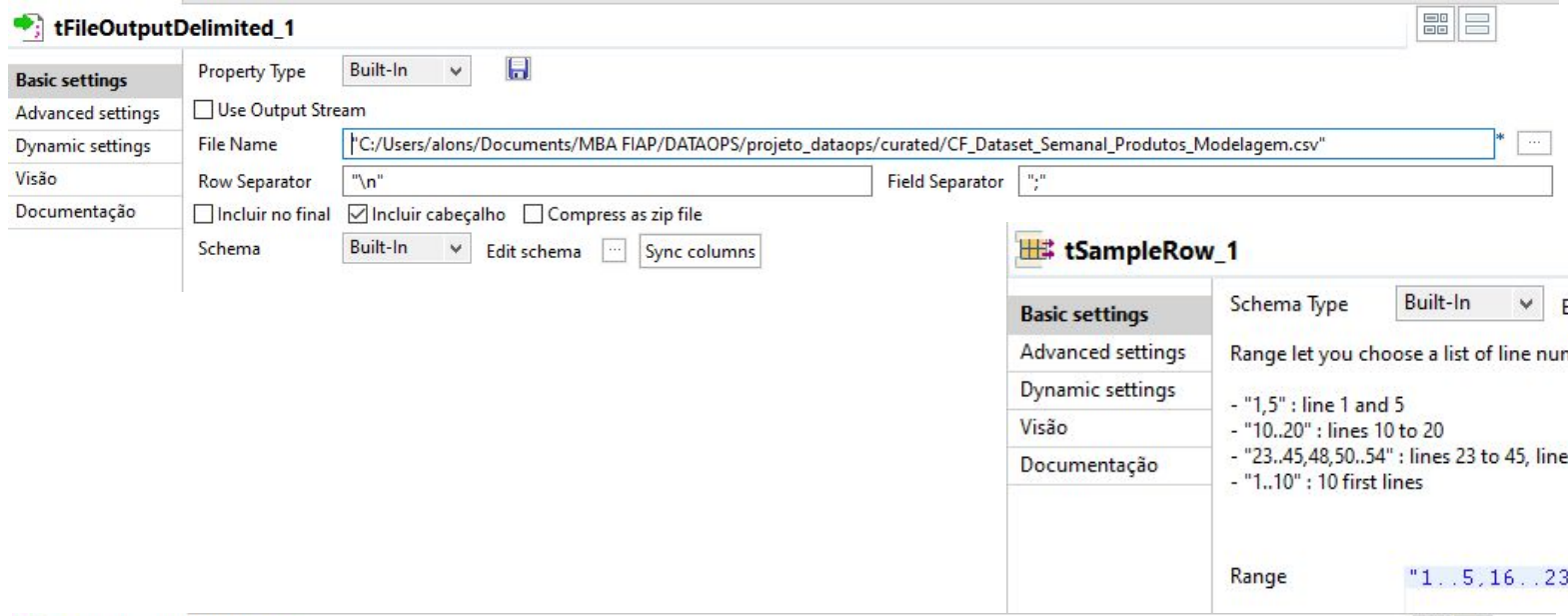
6 - Alteração no Dataset 2

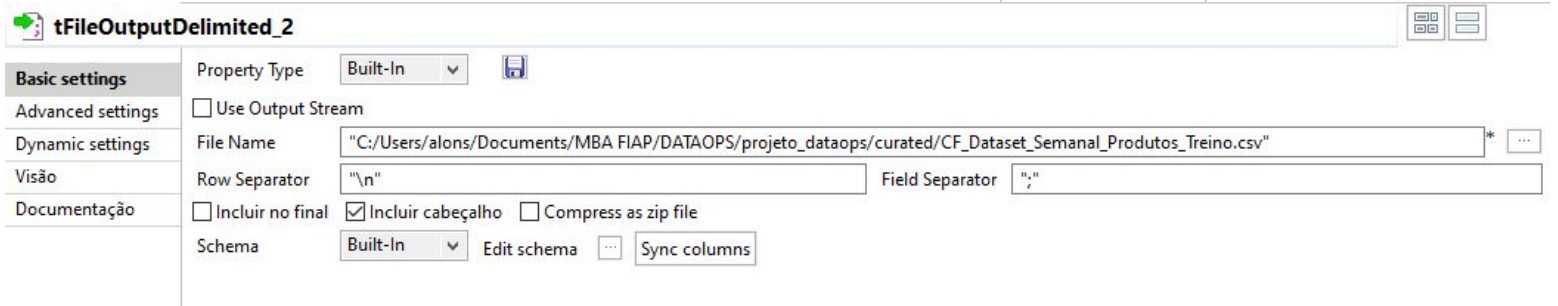


6 - Alteração no Dataset 2



6 - Alteração no Dataset 2





6 - Alteração no Dataset 2 - Resultado

Dataset Para Modelagem

	A	B	C	D	E	F	G	H
1	ProductName	ProductColor	ProductSize	StandardCost	ListPrice	SalesPerson	CountryRegion	OrderDate
2	ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-workslinda3	United States	01/06/2008
3	Rear Brakes	Silver		47.286	106.5	adventure-worksjae0	United Kingdom	01/06/2008
4	Hydration Pack - 70 oz.	Silver	70	205.663	54.99	adventure-worksjae0	United Kingdom	01/06/2008
5	Sport-100 Helmet, Red	Red		130.863	34.99	adventure-workslinda3	United States	01/06/2008
6	Sport-100 Helmet, Black	Black		130.863	34.99	adventure-workslinda3	United States	01/06/2008
7	Sport-100 Helmet, Black	Black		130.863	34.99	adventure-worksjae0	United Kingdom	01/06/2008
8	HL Touring Frame - Yellow, 60	Yellow	60	6.017.437	1003.91	adventure-workslinda3	United States	01/06/2008
9	ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-worksjae0	United Kingdom	01/06/2008
10	Racing Socks, M	White	M	33.623	8.99	adventure-workslinda3	United States	01/06/2008
11	Front Brakes	Silver		47.286	106.5	adventure-worksshu0	United States	01/06/2008
12	Women's Mountain Shorts, S	Black	S	261.763	69.99	adventure-worksjae0	United Kingdom	01/06/2008
13	LL Mountain Frame - Silver, 52	Silver	52	1.445.938	264.05	adventure-worksjae0	United Kingdom	01/06/2008
14	Chain	Silver		89.866	20.24	adventure-worksshu0	United States	01/06/2008
15	Mountain-400-W Silver, 40	Silver	40	4.197.784	769.49	adventure-worksjae0	United Kingdom	01/06/2008
16	Front Derailleur	Silver		406.216	91.49	adventure-worksshu0	United States	01/06/2008
17	AWC Logo Cap	Multi		69.223	8.99	adventure-workslinda3	United States	01/06/2008
18	LL Road Pedal	Silver/Black		179.776	40.49	adventure-workslinda3	United States	01/06/2008
19	HL Road Frame - Red, 62	Red	62	8.686.342	1431.5	adventure-worksjae0	United Kingdom	01/06/2008
20	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksjae0	United Kingdom	01/06/2008
21	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksshu0	United States	01/06/2008
22	Front Brakes	Silver		47.286	106.5	adventure-worksshu0	United States	01/06/2008
23	Touring-3000 Yellow, 44	Yellow	44	4.614.448	742.35	adventure-worksjae0	United Kingdom	01/06/2008
24	HL Mountain Pedal	Silver/Black		359.596	80.99	adventure-worksshu0	United States	01/06/2008

CF_Dataset_Semanal_Produtos_Mod

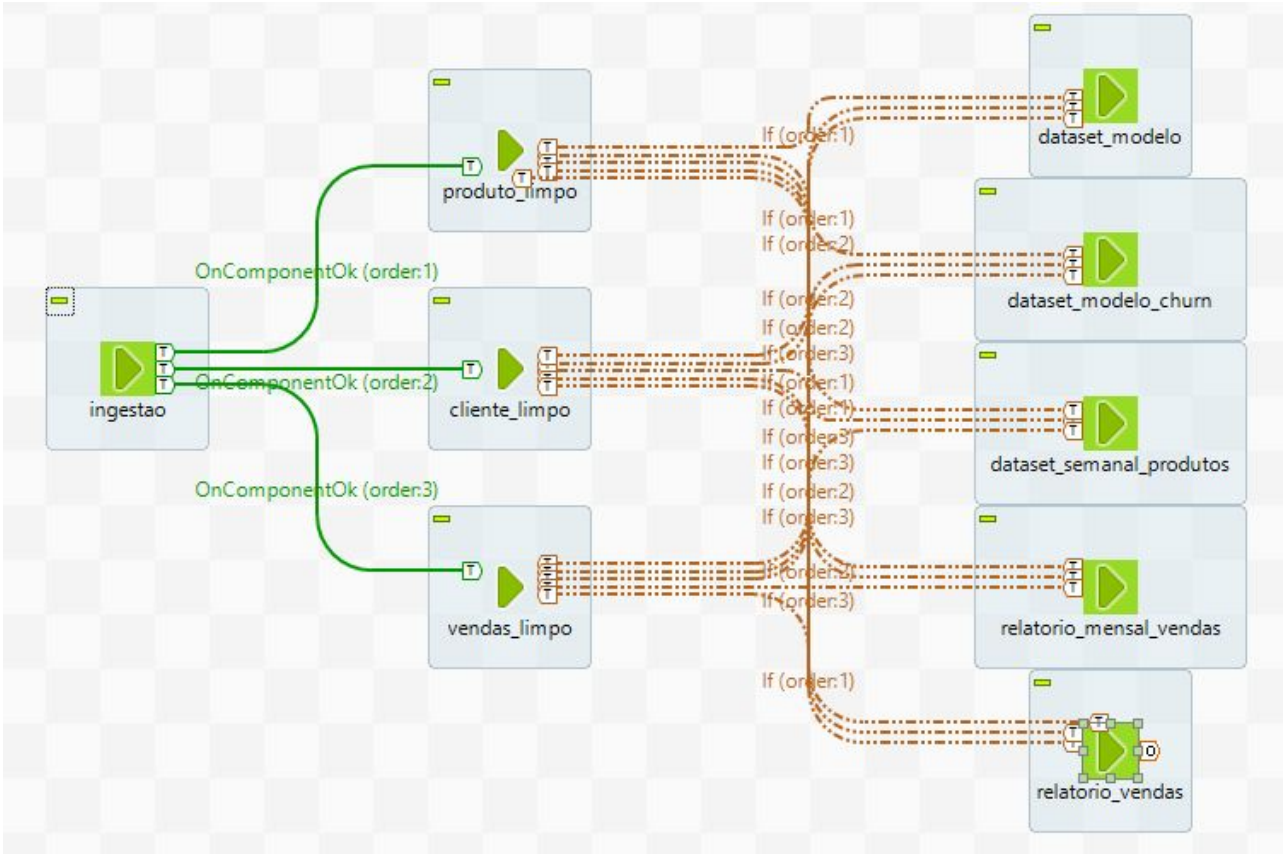
Dataset Para Treinamento

	A	B	C	D	E	F	G	H
1	ProductName	ProductColor	ProductSize	StandardCost	ListPrice	SalesPerson	CountryRegion	OrderDate
2	ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-workslinda3	United States	01/06/2008
3	Rear Brakes	Silver		47.286	106.5	adventure-worksjae0	United Kingdom	01/06/2008
4	Hydration Pack - 70 oz.	Silver	70	205.663	54.99	adventure-worksjae0	United Kingdom	01/06/2008
5	Sport-100 Helmet, Red	Red		130.863	34.99	adventure-workslinda3	United States	01/06/2008
6	Sport-100 Helmet, Black	Black		130.863	34.99	adventure-workslinda3	United States	01/06/2008
7	AWC Logo Cap	Multi		69.223	8.99	adventure-workslinda3	United States	01/06/2008
8	LL Road Pedal	Silver/Black		179.776	40.49	adventure-workslinda3	United States	01/06/2008
9	HL Road Frame - Red, 62	Red	62	8.686.342	1431.5	adventure-worksjae0	United Kingdom	01/06/2008
10	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksjae0	United Kingdom	01/06/2008
11	Classic Vest, S	Blue	S	23.749	63.5	adventure-worksshu0	United States	01/06/2008
12	Front Brakes	Silver		47.286	106.5	adventure-worksshu0	United States	01/06/2008
13	Touring-3000 Yellow, 44	Yellow	44	4.614.448	742.35	adventure-worksjae0	United Kingdom	01/06/2008
14	HL Mountain Pedal	Silver/Black		359.596	80.99	adventure-worksshu0	United States	01/06/2008
15	HL Touring Seat/Saddle			233.722	52.64	adventure-worksjae0	United Kingdom	01/06/2008
16								
17								
18								
19								
20								
21								
22								
23								
24								

CF_Dataset_Semanal_Produtos_Tre

Orquestração do Projeto - Talend

Nesse job, são relacionadas as dependências entre os jobs responsáveis por alimentar as camadas “raw”, “harmonized” e “curated”.



Orquestração do Projeto - DAG 1 Airflow

```
# Importando as bibliotecas que vamos utilizar
from airflow import DAG
from datetime import datetime, timedelta
from airflow.operators.bash_operator import BashOperator
# definição de argumentos básicos
default_args = {
'owner': 'GRUPO4',
'depends_on_past': False,
'start_date': datetime(2021, 5, 2),
'retries': 0
}
# Nomeando a DAG e definindo quando ela vai ser executada - diariamente
with DAG(
'Jobs-data-ops-diario',
schedule_interval='@daily',
catchup=False,
default_args=default_args
) as dag:
# Definindo as tarefas que a DAG vai executar, nesse caso a execução de dois programas Python, chamando sua execução por comandos bash
# O operador Bash, também pode ser utilizado para executar jobs Talend via Sh
    t1 = BashOperator(
    task_id='CamadaRawIngestao',
    bash_command="""
    cd $AIRFLOW_HOME/dags/BuildsJobTalend/Raw/ingestao/ingestao_run.sh
    """)
    t2 = BashOperator(
    task_id='CamadaHarmonizedCliente',
    bash_command="""
    cd $AIRFLOW_HOME/dags/BuildsJobTalend/Harmonized/cliente_limpo/cliente_limpo_run.sh
    """)
        t3 = BashOperator(
    task_id='CamadaHarmonizedProduto',
    bash_command="""
    cd $AIRFLOW_HOME/dags/BuildsJobTalend/Harmonized/produto_limpo/produto_limpo_run.sh
    """)
        t4 = BashOperator(
    task_id='CamadaHarmonizedVendas',
    bash_command="""
    cd $AIRFLOW_HOME/dags/BuildsJobTalend/Harmonized/vendas_limpo;vendas_limpo_run.sh
    """)
        t5 = BashOperator(
    task_id='CamadaCuratedRelatorioVendas',
    bash_command="""
    cd $AIRFLOW_HOME/dags/BuildsJobTalend/Curated/relatorio_vendas/relatorio_vendas_run.sh
    """)
# Definindo o padrão de execução:
t1 >> t2 >> t3 >> t4 >> t5
```

Orquestração do Projeto - DAG 2 Airflow

27 lines (27 sloc) | 956 Bytes

```
# Importando as bibliotecas que vamos utilizar
from airflow import DAG
from datetime import datetime, timedelta
from airflow.operators.bash_operator import BashOperator
# definição de argumentos básicos
default_args = {
'owner': 'GRUPO4',
'depends_on_past': False,
'start_date': datetime(2021, 5, 2),
'retries': 0
}
# Nomeando a DAG e definindo quando ela vai ser executada - semanalmente
with DAG(
'Jobs-data-ops-semanal',
schedule_interval='@weekly',
catchup=False,
default_args=default_args
) as dag:
# Definindo as tarefas que a DAG vai executar, nesse caso a execução de dois programas Python, chamando sua execução por comandos bash
# O operador Bash, também pode ser utilizado para executar jobs Talend via Sh
    t1 = BashOperator(
    task_id='CamadaCurated_dataset_produtos',
    bash_command="""
    cd $AIRFLOW_HOME/dags/dataops/curated/dataset_semanal_produtos/dataset_produtos_run.sh
    """)
# Definindo o padrão de execução:
t1
```

FIAP

MBA+

Engenharia de Dados
DataOps

Aline	336704
Felipe	337491
Heraldo	338426
Stephany	337136