

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

The screenshot shows the GitHub repository page for 'linecoqueto / dataops_fiap'. The repository has 5 branches and 0 tags. The main branch is selected. There are 46 commits in the last pull request. The commits are listed below:

Commit	Message	Time
AirflowDags	Scripts Dags - versao 1	9d82dab 6 minutes ago
BuildsJobTalend	Builds Jobs Talend - versao 1	16 minutes ago
JobsTalend	Jobs Talend - versao 1	2 hours ago
documentacaoHTML	adicionando a documentacao v.1	21 hours ago
Assessment banco de dados.pdf	Assessment BD AdventureWorks	24 days ago
Benefícios do GitFlow.pdf	Benefícios gitflow versao final	2 months ago
README.md	Update README.md	2 months ago

The README.md file contains the following text:

```
Repositório utilizado para os exercícios/trabalhos da matéria de DataOps do MBA FIAP.  
ALINE MAYARA COQUETO LIMA - 336704  
FELIPE LEAL COSTA - 337491  
HERALDO ARAUJO DA SILVA - 338426  
STEPHANY DE CAMILO E ALONSO - 337136
```

The repository has 1 star, 1 fork, and 0 issues. It also has 0 pull requests, 0 actions, 0 projects, 0 wiki pages, and 0 security vulnerabilities.

About: No description, website, or topics provided.

Readme: Readme

Releases: No releases published

Packages: No packages published

Contributors: 3 contributors: linecoqueto (Aline Coqueto), heraldoaraujo (Heraldo Araujo), and alinecoquetolima.

Languages: Languages used: Java (61.2%), HTML (37.9%), CSS (0.2%), Shell (0.2%), PowerShell (0.2%), and Batchfile (0.2%).

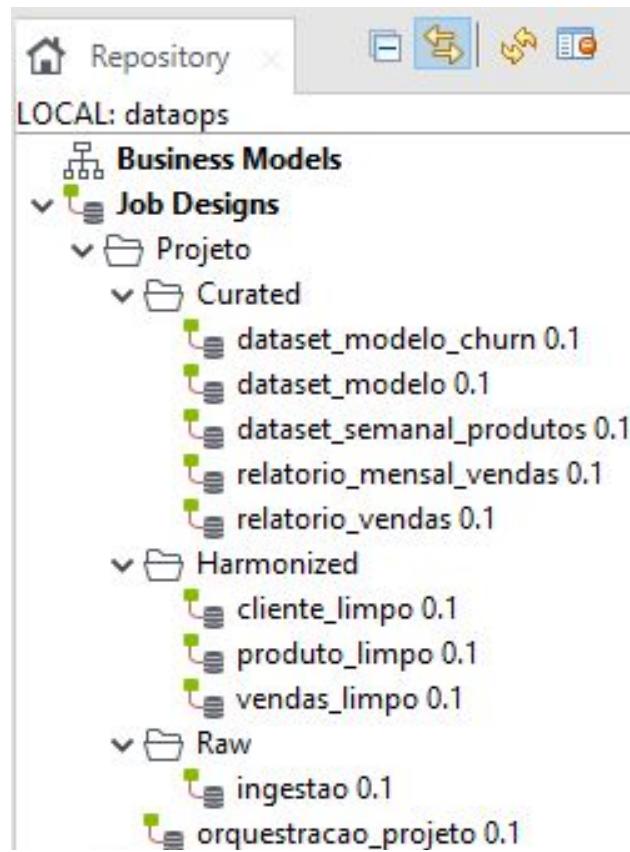
Arquitetura de Dados

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

Database	Microsoft SQL Server	Apply
Property Type	Built-In	
JDBC Provider	Open source JTDS	
Host	"sqlservercentralpublic.database.windows.net"	Porta "1433"
Database	"AdventureWorks"	Schema "SalesLT"
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

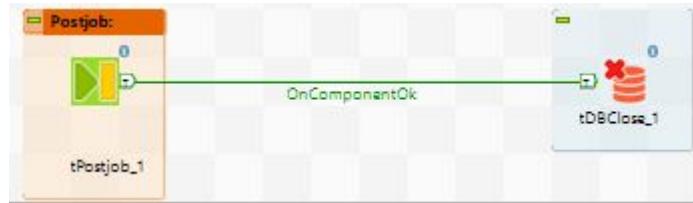
The screenshot shows a Talend job configuration. On the left, a flow diagram has a green arrow pointing from a component labeled "Consulta_SalesOrderDetail" to a component labeled "tFileOutputDelimited_2". The "Consulta_SalesOrderDetail" component has an orange arrow pointing to it from a yellow warning icon. The "tFileOutputDelimited_2" component has a green arrow pointing to it from a blue question mark icon. The main window displays two configuration panels:

- Consulta_SalesOrderDetail(tDBInput_2)(Microsoft SQL Server)**:
 - Basic settings**: Database set to Microsoft SQL Server, Property Type set to Built-In, JDBC Provider set to Open source JTDs.
 - Query**: Contains the following SQL 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**: Property Type set to Built-In.
 - File Name**: Set to "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/raw/SalesOrderDetail.csv".
 - CSV Row Separator**: Set to LF("\n").
 - Field Separator**: Set to ";".
 - Schema**: Set to Built-In.

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

Transformação Camada Raw - Ingestão



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

The screenshot shows the configuration dialog for the tDBClose_1 component. The title bar says 'tDBClose_1 (Microsoft SQL Server)'. On the left is a sidebar with tabs: 'Basic settings' (selected), 'Advanced settings', 'Dynamic settings', 'Visão', and 'Documentação'. The main area has two dropdown menus: 'Database' set to 'Microsoft SQL Server' and 'Lista de componentes' set to 'tDBConnection_1'. There is also an 'Apply' button.

Transformação Camada Raw - Resultados

« Documentos > MBA FIAP > DATAOPS > projeto_dataops > raw

Pesquisar raw

Nome	Data de modificação	Tipo	Tamanho
Address	02/05/2021 10:03	Arquivo de Valore...	66 KB
Customer	02/05/2021 10:03	Arquivo de Valore...	170 KB
Product	02/05/2021 10:03	Arquivo de Valore...	83 KB
ProductCategory	02/05/2021 10:03	Arquivo de Valore...	4 KB
ProductModel	02/05/2021 10:03	Arquivo de Valore...	27 KB
SalesOrderDetail	02/05/2021 10:03	Arquivo de Valore...	60 KB
SalesOrderHeader	02/05/2021 10:03	Arquivo de Valore...	9 KB

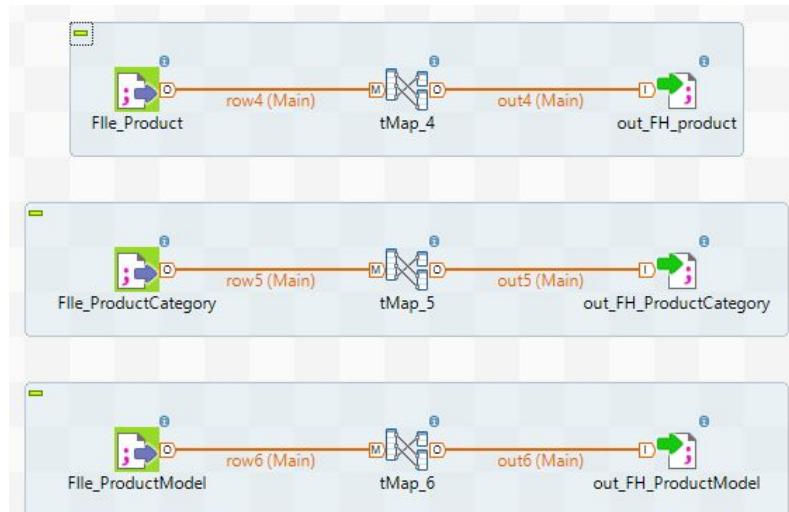
A	B	C	D	E	F	G	H	I	J	K	L	
1	ProductID	ProductNumber	Name	Color	StandardCost	ListPrice	Size	Weight	ProductCategoryID	ProductModelID	SellStartDate	SellEndDate
2	680	FR-R92B-58	HL Road Frame - Black, 58	Black	10.593,100	14.315,000	58	1016,04	18	6	Sat Jun 01 00:00:00 BRT 2002	
3	706	FR-R92R-58	HL Road Frame - 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 BRT 2005
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 BRT 2005
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	UJ-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	UJ-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	UJ-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	UJ-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-R38B-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 BRT 2005
23	726	FR-R38B-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 BRT 2005
24	727	FR-R38B-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 BRT 2005

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.

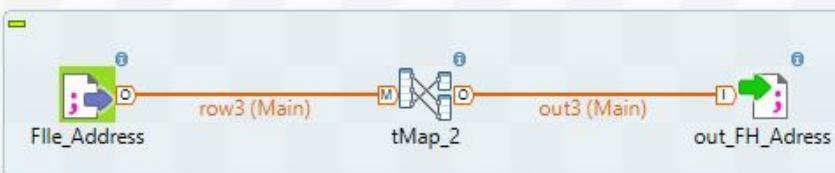
Aqui, os dados são formatados e é feita a seleção das colunas que irão compor os dados dessa camada.



Entidade Produto

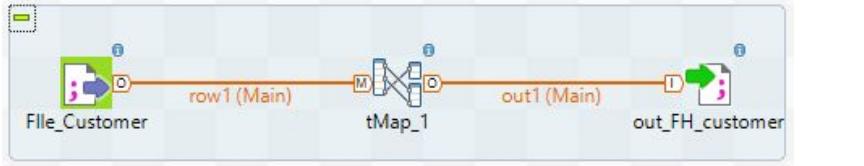


Entidade Vendas



Entidade Cliente

Transformação Camada Harmonized



File_Customer(FileInputDelimited_1)

Basic settings

Property Type: Built-In
Schema: Repository - DELIM:File_customer - metadata
"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."
File Name/Input Stream: C:/Users/alonso/Documents/MBA FIAP/DATAOPS/projeto_dataops/raw/Customer.csv
CSV Row Separator: LF("\n")
Field Separator: ";"
 Opções CSV
Escape char: "\\"
Text enclosure: "\"
Cabeçalho: 1
Rodapé: 0
Limit:
 Skip empty rows
 Uncompress as zip file
 Die on error

out_FH_customer(tFileOutputDelimited_1)

Basic settings

Property Type: Built-In
 Use Output Stream
File Name: C:/Users/alonso/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_customer.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

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.

Talend Open Studio for Big Data - tMap - tMap_1

row1

Column	Type	Key	N.	Date Pa...	Le...	Pre...	D...	Co...
CustomerID	Integer			2	0			
Title	String			3	0			
Suffix	String			3	0			
CompanyName	String			34	0			
SalesPerson	String			23	0			
EmailAddress	String			32	0			
PasswordHash	String			44	0			
PasswordSalt	String			8	0			
rowguid	String			36	0			
ModifiedDate	Date			"dd-M...	10	0		

out1

Column	Type	Key	N.	Date Pa...	Le...	Pre...	D...	Co...
CustomerID	Integer			2	0			
Title	String			3	0			
Suffix	String			3	0			
CompanyName	String			34	0			
SalesPerson	String			23	0			
EmailAddress	String			32	0			

Schema editor Expression editor

row1

Column	Key	Type	Check	N.	Date Pa...	Le...	Pre...	D...	Co...
CustomerID		Int...			2	0			
Title		Stri...			3	0			
Suffix		Stri...			3	0			
CompanyName		Stri...			34	0			
SalesPerson		Stri...			23	0			
EmailAddress		Stri...			32	0			
PasswordHash		Stri...			44	0			
PasswordSalt		Stri...			8	0			
rowguid		Stri...			36	0			
ModifiedDate		Date			"dd-M...	10	0		

out1

Column	Key	Type	Check	N.	Date Pa...	Le...	Pre...	D...	Co...
CustomerID		Integer			2	0			
Title		String			3	0			
Suffix		String			3	0			
CompanyName		String			34	0			
SalesPerson		String			23	0			
EmailAddress		String			32	0			

Apply OK Cancel

Transformação Camada Harmonized - Resultados

« Documentos > MBA FIAP > DATAOPS > projeto_dataops > harmonized

Pesquisar harmon...

Nome	Data de modificação	Tipo	Tamanho
FH_Adress	01/05/2021 20:22	Arquivo de Valore...	35 KB
FH_customer	02/05/2021 10:08	Arquivo de Valore...	78 KB
FH_product	01/05/2021 20:22	Arquivo de Valore...	40 KB
FH_ProductCategory	01/05/2021 20:22	Arquivo de Valore...	1 KB
FH_ProductModel	01/05/2021 20:22	Arquivo de Valore...	4 KB
FH_SalesOrderDetail	01/05/2021 20:22	Arquivo de Valore...	29 KB
FH_SalesOrderHeader	01/05/2021 20:22	Arquivo de Valore...	7 KB

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	S071774	PO348186287
3	71776	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071776	PO19952192051
4	71780	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071780	PO19604173239
5	71782	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071782	PO19372114749
6	71783	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071783	PO19285135919
7	71784	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071784	PO19285135919
8	71796	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071796	PO17052159664
9	71797	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071797	PO16501134889
10	71815	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071815	PO13021155785
11	71816	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071816	PO12992180445
12	71831	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071831	PO10295111084
13	71832	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071832	PO10353140756
14	71845	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071845	PO2697119362
15	71846	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071846	PO2378131604
16	71856	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071856	PO16530177647
17	71858	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071858	PO16153112278
18	71863	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071863	PO16124166561
19	71867	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071867	PO13050111529
20	71885	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071885	PO6119130779
21	71895	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071895	PO3770176273
22	71897	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071897	PO4785152479
23	71898	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071898	PO5713190501
24	71899	2	01/06/2008	13/06/2008	08/06/2008	5	false	S071899	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.

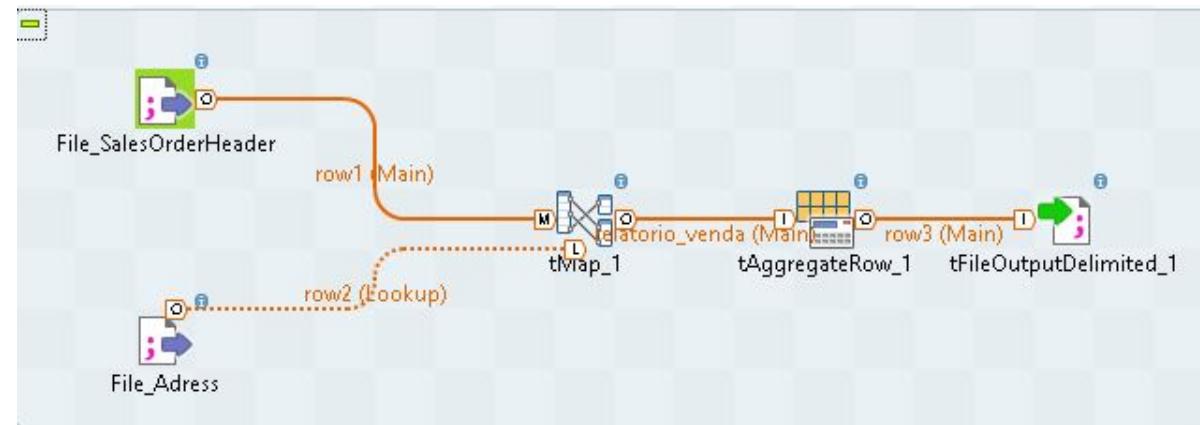
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings

Property Type: Built-In

Schema: Built-In

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv"

CSV Row Separator: LF("\n") Field Separator: ";"

Opções CSV Escape char: "" Text enclosure: ""

Cabeçalho: 1 Rodapé: 0 Limit:

Skip empty rows Uncompress as zip file Die on error

File_Adress(tFileInputDelimited_3)

Basic settings

Property Type: Built-In

Schema: Repository DELIM:File_Harmonized_Address - metadata

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_Adress.csv"

CSV Row Separator: LF("\n") Field Separator: ";"

Opções CSV Escape char: "" Text enclosure: ""

Cabeçalho: 1 Rodapé: 0 Limit:

Skip empty rows Uncompress as zip file Die on error

1 - Relatório para área de vendas

Talend Open Studio for Big Data - tMap - tMap_1

row1

Column
SalesOrderID
RevisionNumber
OrderDate
DueDate
ShipDate
Status
OnlineOrderFlag
SalesOrderNumber
PurchaseOrderNumber
AccountNumber
CustomerID
ShipToAddressID
BillToAddressID
ShipMethod
CreditCardApprovalCode
SubTotal
TaxAmt
Freight
TotalDue

Var

relatorio_venda

Expressão	Column
row1.OrderDate	OrderDate
row1.OnlineOrderFlag	OnlineOrderFlag
row1.TotalDue	TotalSold
row2.CountryRegion	CountryRegion

Schema editor

Column	Key	Type	N.	Date Pattern (...)	Length	Precision	Defa...	Comen...
SalesOrderID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	28	0		
DueDate	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		28	0		

Expression editor

Column	Key	Type	N.	Date Pattern (...)	Length	Precision	Defa...	Comen...
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	28	0		
OnlineOrderFlag	<input type="checkbox"/>	Boolean	<input checked="" type="checkbox"/>		5	0		
TotalSold	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		9	2		
CountryRegion	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		13	0		

1 - Relatório para área de vendas

tAggregateRow_1

Basic settings

Schema: Built-In | Edit schema | ... | Sync columns

Grupo por:

Output column	Input column position
OrderDate	OrderDate
OnlineOrderFlag	OnlineOrderFlag
CountryRegion	CountryRegion

Operations:

Output column	Função	Input column position	<input type="checkbox"/> Ignore null values
TotalDue	sum	TotalSold	<input checked="" type="checkbox"/>

tFileOutputDelimited_1

Basic settings

Property Type: Built-In |

Advanced settings: Use Output Stream

Dynamic settings: File Name: "C:/Users/alonso/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Relatorio_Vendas.csv" | ...

Visão: Use OS line separator as row separator when CSV Row Separator is set to CR,LF or CRLF.

Documentação: CSV Row Separator: LF("\n") | Field Separator: ";"

Incluir no final | Incluir cabecalho | Compress as zip file

Schema: Built-In | Edit schema | ... | Sync columns

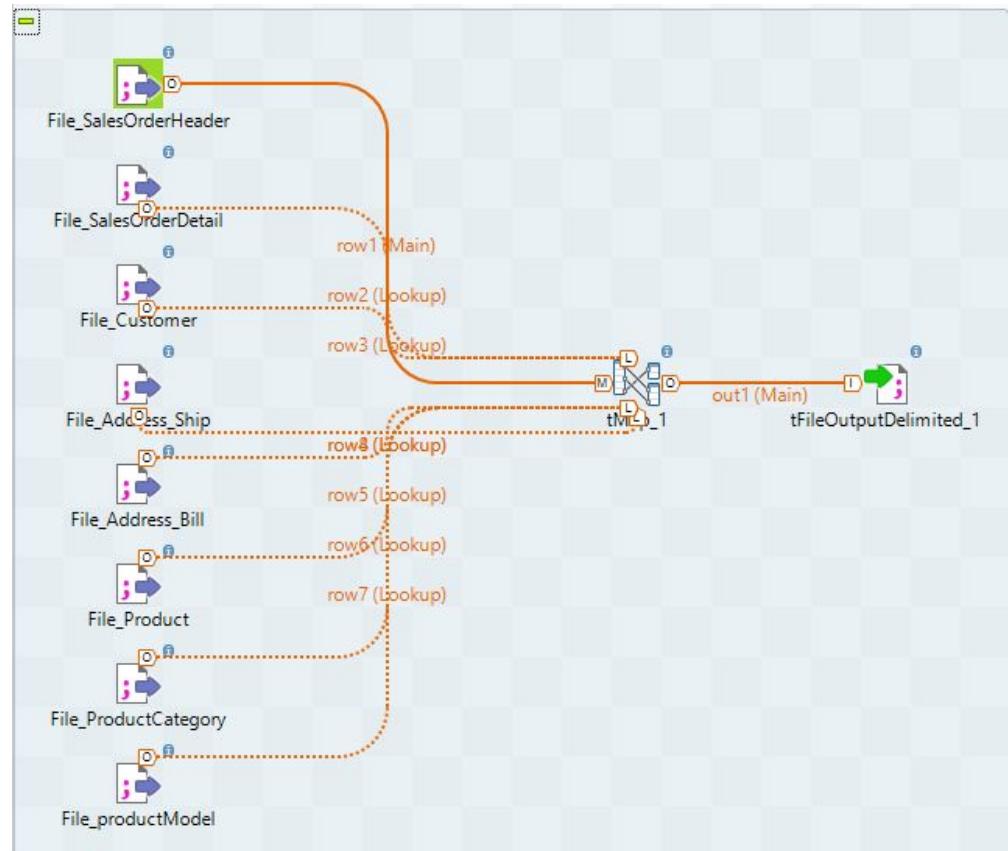
2 - Dataset para exploração - ciência de dados

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings	Property Type <input type="button" value="Repository"/> DELIM:File_Harmonized_SalesOrderHeader <input type="button"/>
Advanced settings	Schema <input type="button" value="Repository"/> DELIM:File_Harmonized_SalesOrderHeader* <input type="button"/> Edit schema <input type="button"/>
Dynamic settings	"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."
Visão	File Name/Input Stream "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv" <input type="button"/>
Documentação	CSV Row Separator <input \n")"="" type="button" value="LF("/> Field Separator ";" <input checked="" type="checkbox"/> Opções CSV Escape char "\\" Text enclosure "\" Cabeçalho <input type="button" value="1"/> Rodapé <input type="button" value="0"/> Limit <input type="button"/> <input type="checkbox"/> Skip empty rows <input type="checkbox"/> Uncompress as zip file <input type="checkbox"/> Die on error

tFileOutputDelimited_1

Basic settings	Property Type <input type="button" value="Built-In"/> <input type="button"/>
Advanced settings	<input type="checkbox"/> Use Output Stream
Dynamic settings	File Name "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Dataset_DataScientists.csv" * <input type="button"/>
Visão	<input checked="" type="checkbox"/> Use OS line separator as row separator when CSV Row Separator is set to CR,LF or CRLF.
Documentação	CSV Row Separator <input \n")"="" type="button" value="LF("/> Field Separator ";" <input type="checkbox"/> Incluir no final <input checked="" type="checkbox"/> Incluir cabeçalho <input type="checkbox"/> Compress as zip file Schema <input type="button" value="Built-In"/> <input type="button"/> Edit schema <input type="button"/> Sync columns

2 - Dataset para exploração - ciência de dados

Talend Open Studio for Big Data - tMap - tMap_1

The screenshot shows the Talend Open Studio for Big Data interface, specifically the tMap component. It consists of three main panels: row1, Var, and out1.

row1: This panel lists the input columns from the SalesOrder table. The columns are: SalesOrderID, RevisionNumber, OrderDate, DueDate, ShipDate, Status, OnlineOrderFlag, SalesOrderNumber, PurchaseOrderNumber, AccountNumber, CustomerID, ShipToAddressID, BillToAddressID, ShipMethod, CreditCardApprovalCode, SubTotal, TaxAmt, Freight, and TotalDue. The "Freight" column is highlighted in yellow.

Var: This panel contains a single entry: "Var".

out1: This panel lists the output columns. The columns are: SalesOrderID, RevisionNumber, OrderDate, DueDate, ShipDate, Status, OnlineOrderFlag, CustomerID, ShipCity, ShipStateProvince, ShipCountryRegion, ShipPostalCode, BillCity, BillStateProvince, BillCountryRegion, BillPostalCode, ShipMethod, CreditCardApprovalCode, SubTotal, and TaxAmt. The "Freight" column is listed under "out1" but is not mapped to any output column.

Schema editor: Below the main panels, there are two schema editors for row1 and out1. Both editors show the mapping between columns and their properties (Key, Type, N., Date Pattern, Length, Precision, Default, Come...).

Column	Key	Type	N.	Date Pattern	Length	Precision	Default	Come...
SalesOrderID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
DueDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
ShipDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		

Column	Key	Type	N.	Date Pattern	Length	Precision	Default	Come...
SalesOrderID	<input checked="" type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
DueDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
ShipDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		

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	71901	2	01/06/2008	13/06/2008	08/06/2008	5	false	29930	Fullerton	California	United States	92821	Fullerton	Calif

3 - Alteração no Relatório 1

Nesse job, são coletados dados das tabelas SalesOrderHeader, Address e Customer.

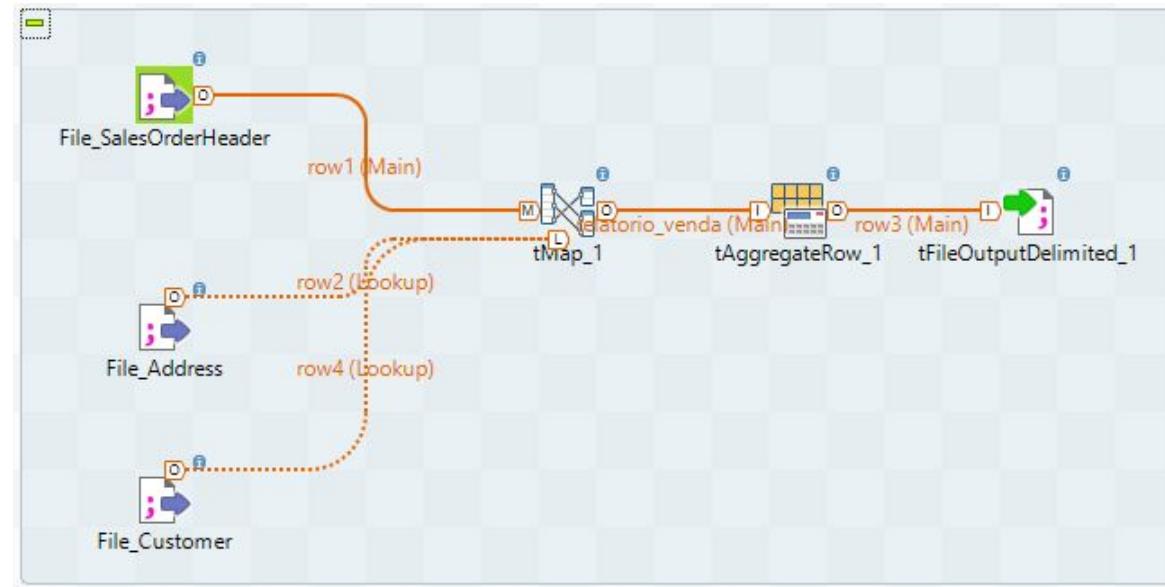
No step tMap são feitos os joins entre as tabelas e são selecionadas as colunas necessárias no relatório e são feitas as transformações necessárias nos nomes de colunas.

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings

Property Type: Built-In

Schema: Repository **DELIM:File_Harmonized_SalesOrderHeader** * Edit schema

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv" *

CSV Row Separator: LF("\n") Field Separator: ";"

Opções CSV Escape char: "" Text enclosure: "\""

Cabeçalho: 1 Rodapé: 0 Limit:

Skip empty rows Uncompress as zip file Die on error

File_Customer(tFileInputDelimited_2)

Basic settings

Property Type: Built-In

Schema: Repository **DELIM:File_Harmonized_Customer - metad** * Edit schema

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_customer.csv" *

CSV Row Separator: LF("\n") Field Separator: ";"

Opções CSV Escape char: "" Text enclosure: "\""

Cabeçalho: 1 Rodapé: 0 Limit:

Skip empty rows Uncompress as zip file Die on error

File_Address(tFileInputDelimited_3)

Basic settings

Property Type: Built-In

Schema: Repository **DELIM:File_Harmonized_Address - metadat** * Edit schema

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_Adress.csv" *

CSV Row Separator: LF("\n") Field Separator: ";"

Opções CSV Escape char: "" Text enclosure: "\""

Cabeçalho: 1 Rodapé: 0 Limit:

Skip empty rows Uncompress as zip file Die on error

3 - Alteração no Relatório 1

Talend Open Studio for Big Data - tMap - tMap_1

Find :

Var

Auto map!

relatorio_venda

Expressão	Column
row1.OrderDate	OrderDate
row1.OnlineOrderFlag	OnlineOrderFlag
row1.TotalDue	TotalSold
row2.CountryRegion	CountryRegion
row4.SalesPerson	SalesPerson

Schema editor Expression editor

row1	relatorio_venda
Column	Column
SalesOrderID	OrderDate
RevisionNumber	OnlineOrderFlag
OrderDate	TotalSold
DueDate	CountryRegion
ShipDate	SalesPerson
Status	
OnlineOrderFlag	
SalesOrderNumber	
PurchaseOrderNumber	
AccountNumber	
CustomerID	
ShipToAddressID	
BillToAddressID	
ShipMethod	
CreditCardApprovalCode	
SubTotal	
TaxAmt	
Freight	
TotalDue	

row1

Column	Type	Key	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
SalesOrderID	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>		1	0		
OrderDate	Date	<input type="checkbox"/>	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	28	0		
DueDate	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>		28	0		
ShipDate	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>		28	0		

relatorio_venda

Column	Type	Key	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
OrderDate	Date	<input type="checkbox"/>	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	28	0		
OnlineOrderFlag	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>		5	0		
TotalSold	Float	<input type="checkbox"/>	<input checked="" type="checkbox"/>		9	2		
CountryRegion	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>		13	0		

3 - Alteração no Relatório 1

tAggregateRow_1

Basic settings

Schema: Built-In

Advanced settings

Dynamic settings

Visão

Documentação

Grupo por

Output column	Input column position
OrderDate	OrderDate
CountryRegion	CountryRegion
OnlineOrderFlag	OnlineOrderFlag
SalesPerson	SalesPerson

Operations

Output column	Função	Input column position	<input type="checkbox"/> Ignore null values
TotalSold	sum	TotalSold	<input checked="" type="checkbox"/>

tFileOutputDelimited_1

Basic settings

Property Type: Built-In

Advanced settings

Dynamic settings

Visão

Documentação

Use Output Stream

File Name: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Relatorio_Vendas.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

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.

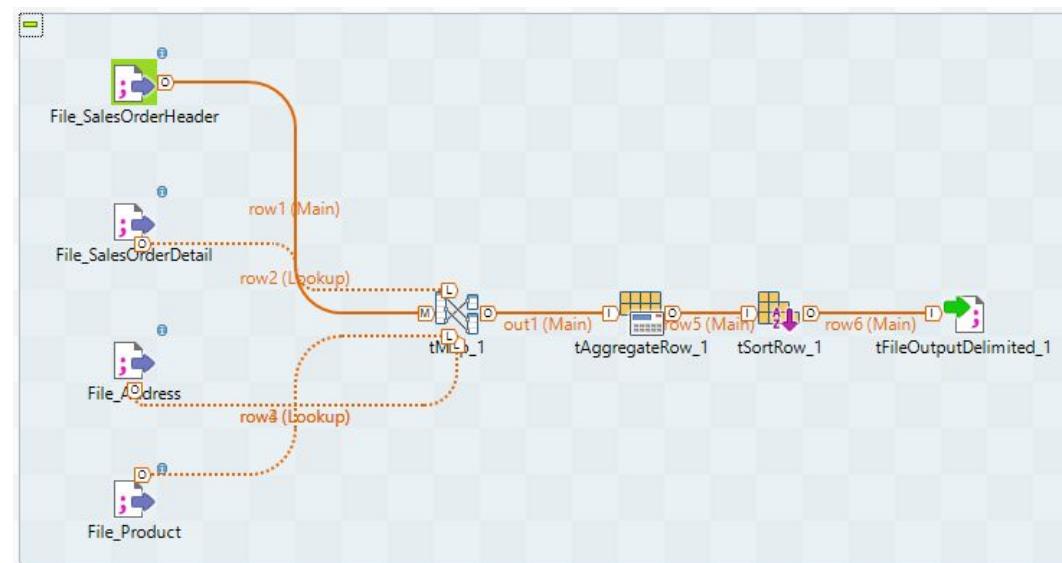
No step tAggregateRow, são feitas as agregações dos dados pelos campos de nome do produto, mês de venda e região.

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings	Property Type <input type="button" value="Built-In"/>	<input type="button" value=""/>
Advanced settings	Schema <input type="button" value="Repository"/>	DELIM:File_Harmonized_SalesOrderHeader * <input type="button" value="..."/> Edit schema <input type="button" value="..."/>
Dynamic settings	"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."	
Visão		
Documentação	File Name/Input Stream "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv" * <input type="button" value="..."/>	
	CSV Row Separator <input \n")"="" type="button" value="LF("/>	Field Separator <input type="button" value=";"/>
	<input checked="" type="checkbox"/> Opções CSV	Escape char <input "="" type="button" value=""/> Text enclosure <input "="" type="button" value="\"/> "
	Cabeçalho <input type="button" value="1"/>	Rodapé <input type="button" value="0"/> Limit <input type="button" value=""/>
	<input type="checkbox"/> Skip empty rows	<input type="checkbox"/> Uncompress as zip file
	<input type="checkbox"/> Die on error	

File_SalesOrderDetail(tFileInputDelimited_3)

Basic settings	Property Type <input type="button" value="Built-In"/>	<input type="button" value=""/>
Advanced settings	Schema <input type="button" value="Repository"/>	DELIM:File_Harmonized_SalesOrderDetail - * <input type="button" value="..."/> Edit schema <input type="button" value="..."/>
Dynamic settings	"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."	
Visão		
Documentação	File Name/Input Stream "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderDetail.csv" * <input type="button" value="..."/>	
	CSV Row Separator <input \n")"="" type="button" value="LF("/>	Field Separator <input type="button" value=";"/>
	<input checked="" type="checkbox"/> Opções CSV	Escape char <input "="" type="button" value=""/> Text enclosure <input "="" type="button" value="\"/> "
	Cabeçalho <input type="button" value="1"/>	Rodapé <input type="button" value="0"/> Limit <input type="button" value=""/>
	<input type="checkbox"/> Skip empty rows	<input type="checkbox"/> Uncompress as zip file
	<input type="checkbox"/> Die on error	

4 - Relatório de produtos vendidos

The screenshot shows the Talend Open Studio for Big Data interface with the tMap component open. The workspace is divided into several panels:

- Row Input (row1):** Contains a list of columns from the input dataset, including SalesOrderID, RevisionNumber, OrderDate, DueDate, ShipDate, Status, OnlineOrderFlag, SalesOrderNumber, PurchaseOrderNumber, AccountNumber, CustomerID, ShipToAddressID, BillToAddressID, ShipMethod, CreditCardApprovalCode, SubTotal, TaxAmt, Freight, and TotalDue.
- Variables (Var):** A central panel showing a variable named "Var".
- Row Output (out1):** Maps the input columns to output columns. The mappings are:
 - row4.Name → Column: ProductName
 - TalendDate.getMonth("MONTH", r...) → Column: OrderMonth
 - row3.CountryRegion → Column: CountryRegion
 - row2.OrderQty → Column: OrderQty
 - row1.TotalDue → Column: TotalSold
- Schema editor:** Shows the schema for the row1 and out1 datasets, detailing column names, types, keys, and patterns.
- Expression editor:** Located below the schema editor, it is used for defining complex mappings.

4 - Relatório de produtos vendidos

tAggregateRow_1

Basic settings Schema Type Built-In Edit schema Sync columns

Advanced settings

Dynamic settings

Visão

Documentação

Grupo por

Output column	Input column position
ProductName	ProductName
OrderMonth	OrderMonth
CountryRegion	CountryRegion

Operations

Output column	Função	Input column position	Ignore null values
OrderQty	sum	OrderQty	<input type="checkbox"/>
TotalSold	sum	TotalSold	<input type="checkbox"/>

tSortRow_1

Basic settings Schema Type Built-In Edit schema Sync columns

Advanced settings

Dynamic settings

Visão

Documentação

Critério

Coluna do esquema	sort num or alpha?	Ordenar asc ou desc?
ProductName	alpha	ascendente
OrderMonth	num	ascendente
OrderQty	num	desc

tFileOutputDelimited_1

Basic settings Property Type Built-In 

Advanced settings Use Output Stream

Dynamic settings

Visão

Documentação

File Name "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Relatorio_Mensal_Produtos.csv" * 

Row Separator "\n" Field Separator ";"

Incluir no final Incluir cabeçalho Compress as zip file

Schema Built-In Edit schema Sync columns

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings	Property Type <input type="button" value="Built-In"/>	<input type="button" value=""/>
Advanced settings	Schema <input type="button" value="Built-In"/>	<input type="button" value="Edit schema"/>
Dynamic settings	"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."	
Visão		
Documentação	File Name/Input Stream "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv" * <input type="button" value="..."/>	
	CSV Row Separator <input \n")"="" type="button" value="LF("/>	Field Separator <input type="button" value=";"/>
	<input checked="" type="checkbox"/> Opções CSV	Escape char <input type="button" value="'''"/>
		Text enclosure <input '''"="" type="button" value="\"/>
	Cabeçalho <input type="button" value="1"/>	Rodapé <input type="button" value="0"/>
	<input type="checkbox"/> Skip empty rows	<input type="checkbox"/> Uncompress as zip file
	<input type="checkbox"/> Die on error	

tFileOutputDelimited_1

Basic settings	Property Type <input type="button" value="Built-In"/>	<input type="button" value=""/>
Advanced settings	<input type="checkbox"/> Use Output Stream	
Dynamic settings	File Name "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Dataset_Semanal_Produtos.csv" * <input type="button" value="..."/>	
Visão	Row Separator <input type="button" value="\n"/>	Field Separator <input type="button" value=";"/>
Documentação	<input type="checkbox"/> Incluir no final	<input checked="" type="checkbox"/> Incluir cabeçalho <input type="checkbox"/> Compress as zip file
	Schema <input type="button" value="Built-In"/>	<input type="button" value="Edit schema"/>
		<input type="button" value="Sync columns"/>

5 - Dataset de produtos vendidos

Talend Open Studio for Big Data - tMap - tMap_1

row1

Column	Key	Type	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
SalesOrderID	<input checked="" type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
DueDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		
ShipDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yy...	10	0		

out1

Column	Key	Type	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
ProductName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		30	0		
ProductColor	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		6	0		
ProductSize	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
StandardCost	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		9	5		
ListPrice	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		7	3		

Schema editor Expression editor

row1

out1

Apply OK Cancel

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-worksshhu0	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-worksshhu0	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-worksshhu0	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-worksshhu0	United States	01/06/2008
22	Front Brakes	Silver		47.286	106.5	adventure-worksshhu0	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-worksshhu0	United States	01/06/2008
25	Classic Vest, M	Blue	M	23.749	63.5	adventure-worksjae0	United Kingdom	01/06/2008

6 - Alteração no Dataset 2

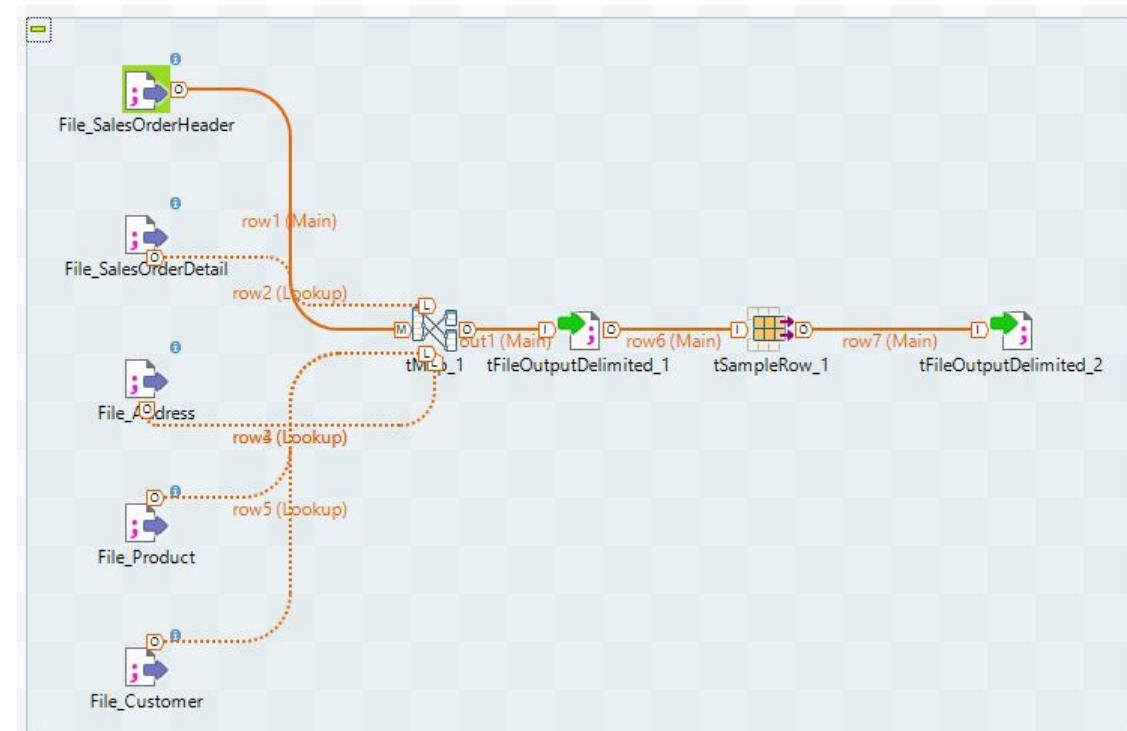
Nesse job, são coletados dados das tabelas das entidades de vendas, vendedores e produto, conforme solicitado.

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

File_SalesOrderHeader(tFileInputDelimited_1)

Basic settings

Property Type: Built-In 

Schema: Built-In  Edit schema 

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderHeader.csv" 

CSV Row Separator: LF("\n")  Field Separator: ";" 

Opções CSV Escape char: ""  Text enclosure: "\" 

Cabeçalho: 1 Rodapé: 0 Limit: 

Skip empty rows Uncompress as zip file Die on error

File_SalesOrderDetail(tFileInputDelimited_2)

Basic settings

Property Type: Built-In 

Schema: Built-In  Edit schema 

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

File Name/Input Stream: "C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/harmonized/FH_SalesOrderDetail.csv" 

CSV Row Separator: LF("\n")  Field Separator: ";" 

Opções CSV Escape char: ""  Text enclosure: "\" 

Cabeçalho: 1 Rodapé: 0 Limit: 

Skip empty rows Uncompress as zip file Die on error

6 - Alteração no Dataset 2

Talend Open Studio for Big Data - tMap - tMap_1

row1

Var

out1

Schema editor Expression editor

row1

Column	Key	Type	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
SalesOrderID	<input checked="" type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
RevisionNumber	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
OrderDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	10	0		
DueDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	10	0		
ShipDate	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	10	0		

out1

Column	Key	Type	N.	Date Pattern ...	Length	Precisi...	Defa...	Come...
ProductName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		30	0		
ProductColor	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		6	0		
ProductSize	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
StandardCost	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		9	5		
ListPrice	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		7	3		

6 - Alteração no Dataset 2

tFileOutputDelimited_1

Basic settings

Property Type Built-In

Use Output Stream

File Name `C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Dataset_Semanal_Produtos_Modelagem.csv`

Row Separator `\n` Field Separator `,`

Incluir no final Incluir cabeçalho Compress as zip file

Schema Built-In

tSampleRow_1

Basic settings

Schema Type Built-In

Advanced settings

Dynamic settings

Visão

Documentação

Range let you choose a list of line numbers and/or a list of ranges.

- "1,5" : line 1 and 5
- "10..20" : lines 10 to 20
- "23..45,48,50..54" : lines 23 to 45, line 48 and lines 50 to 54
- "1..10" : 10 first lines

Range `1..5,16..23,32..37,47..53`

tFileOutputDelimited_2

Basic settings

Property Type Built-In

Use Output Stream

File Name `C:/Users/alons/Documents/MBA FIAP/DATAOPS/projeto_dataops/curated/CF_Dataset_Semanal_Produtos_Treino.csv`

Row Separator `\n` Field Separator `,`

Incluir no final Incluir cabeçalho Compress as zip file

Schema Built-In

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-worksinda3	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-worksinda3	United States	01/06/2008
6 Sport-100 Helmet, Black	Black		130.863	34.99	adventure-worksinda3	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-worksinda3	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-worksinda3	United States	01/06/2008
11 Front Brakes	Silver		47.286	106.5	adventure-workssh0	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-workssh0	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-workssh0	United States	01/06/2008
17 AWC Logo Cap	Multi		69.223	8.99	adventure-worksinda3	United States	01/06/2008
18 LL Road Pedal	Silver/Black		179.776	40.49	adventure-worksinda3	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-workssh0	United States	01/06/2008
22 Front Brakes	Silver		47.286	106.5	adventure-workssh0	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-workssh0	United States	01/06/2008
25 Classic Vest, S	Blue		23.749	63.5	adventure-worksjae0	United States	01/06/2008

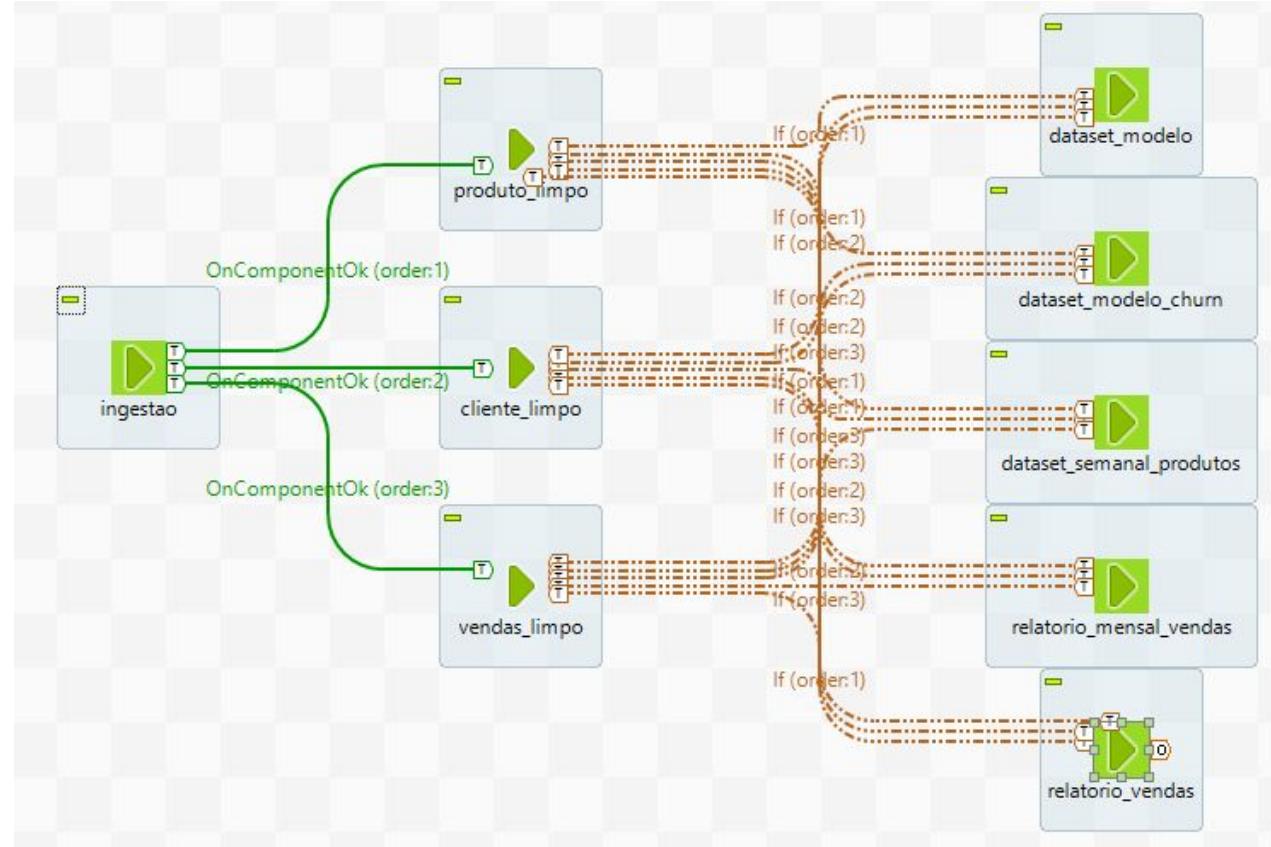
A	B	C	D	E	F	G	H
1 ProductName	ProductColor	ProductSize	StandardCost	ListPrice	SalesPerson	CountryRegion	OrderDate
2 ML Road Frame-W - Yellow, 38	Yellow	38	3.609.428	594.83	adventure-worksinda3	United States	01/06/2008
3 Rear Brakes	Silver		47.286	106.5	adventure-worksjae0	United Kingdom	01/06/2008
4 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-worksinda3	United States	01/06/2008
6 Sport-100 Helmet, Black	Black		130.863	34.99	adventure-worksinda3	United States	01/06/2008
7 AWC Logo Cap	Multi		69.223	8.99	adventure-worksinda3	United States	01/06/2008
8 LL Road Pedal	Silver/Black		179.776	40.49	adventure-worksinda3	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-workssh0	United States	01/06/2008
12 Front Brakes	Silver		47.286	106.5	adventure-workssh0	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-workssh0	United States	01/06/2008
15 HL Touring Seat/Saddle	Silver/Black		233.722	52.64	adventure-worksjae0	United Kingdom	01/06/2008
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							

CF_Dataset_Semanal_Produtos_Tre

Dataset Para Treinamento

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

```
 1 # Importando as bibliotecas que vamos utilizar
 2 from airflow import DAG
 3 from datetime import datetime, timedelta
 4 from airflow.operators.bash_operator import BashOperator
 5 # definição de argumentos básicos
 6 default_args = {
 7     'owner': 'GRUPO4',
 8     'depends_on_past': False,
 9     'start_date': datetime(2021, 5, 2),
10     'retries': 0
11 }
12 # Nomeando a DAG e definindo quando ela vai ser executada - diariamente
13 with DAG(
14     'Jobs-data-ops-diario',
15     schedule_interval='@daily',
16     catchup=False,
17     default_args=default_args
18 ) as dag:
19     # Definindo as tarefas que a DAG vai executar, nesse caso a execução de dois programas Python, chamando sua execução por comandos bash
20     # O operador Bash, também pode ser utilizado para executar jobs Talend via Sh
21     t1 = BashOperator(
22         task_id='CamadaRawIngestao',
23         bash_command="""
24             cd ${AIRFLOW_HOME}/dags/BuildsJobTalend/Raw/ingestao/ingestao_run.sh
25         """)
26     t2 = BashOperator(
27         task_id='CamadaHarmonizedCliente',
28         bash_command="""
29             cd ${AIRFLOW_HOME}/dags/BuildsJobTalend/Harmonized/cliente_limpo/cliente_limpo_run.sh
30         """)
31     t3 = BashOperator(
32         task_id='CamadaHarmonizedProduto',
33         bash_command="""
34             cd ${AIRFLOW_HOME}/dags/BuildsJobTalend/Harmonized/produto_limpo/produto_limpo_run.sh
35         """)
36     t4 = BashOperator(
37         task_id='CamadaHarmonizedVendas',
38         bash_command="""
39             cd ${AIRFLOW_HOME}/dags/BuildsJobTalend/Harmonized/vendas_limpo;vendas_limpo_run.sh
40         """)
41     t5 = BashOperator(
42         task_id='CamadaCuratedRelatorioVendas',
43         bash_command="""
44             cd ${AIRFLOW_HOME}/dags/BuildsJobTalend/Curated/relatorio_vendas/relatorio_vendas_run.sh
45         """)
46     # Definindo o padrão de execução:
47     t1 >> t2 >> t3 >> t4 >> t5
```

Orquestração do Projeto - DAG 2 Airflow

```
27 lines (27 sloc) | 956 Bytes

1  # Importando as bibliotecas que vamos utilizar
2  from airflow import DAG
3  from datetime import datetime, timedelta
4  from airflow.operators.bash_operator import BashOperator
5  # definição de argumentos básicos
6  default_args = {
7      'owner': 'GRUPO04',
8      'depends_on_past': False,
9      'start_date': datetime(2021, 5, 2),
10     'retries': 0
11 }
12 # Nomeando a DAG e definindo quando ela vai ser executada - semanalmente
13 with DAG(
14     'Jobs-data-ops-semanal',
15     schedule_interval='@weekly',
16     catchup=False,
17     default_args=default_args
18 ) as dag:
19     # Definindo as tarefas que a DAG vai executar, nesse caso a execução de dois programas Python, chamando sua execução por comandos bash
20     # O operador Bash, também pode ser utilizado para executar jobs Talend via Sh
21     t1 = BashOperator(
22         task_id='CamadaCurated_dataset_produtos',
23         bash_command="""
24             cd $AIRFLOW_HOME/dags/dataops/curated/dataset_semanal_produtos/dataset_produtos_run.sh
25         """
26     # Definindo o padrão de execução:
27     t1
```

MBA⁺

Engenharia de Dados
DataOps

Aline	336704
Felipe	337491
Heraldo	338426
Stephany	337136