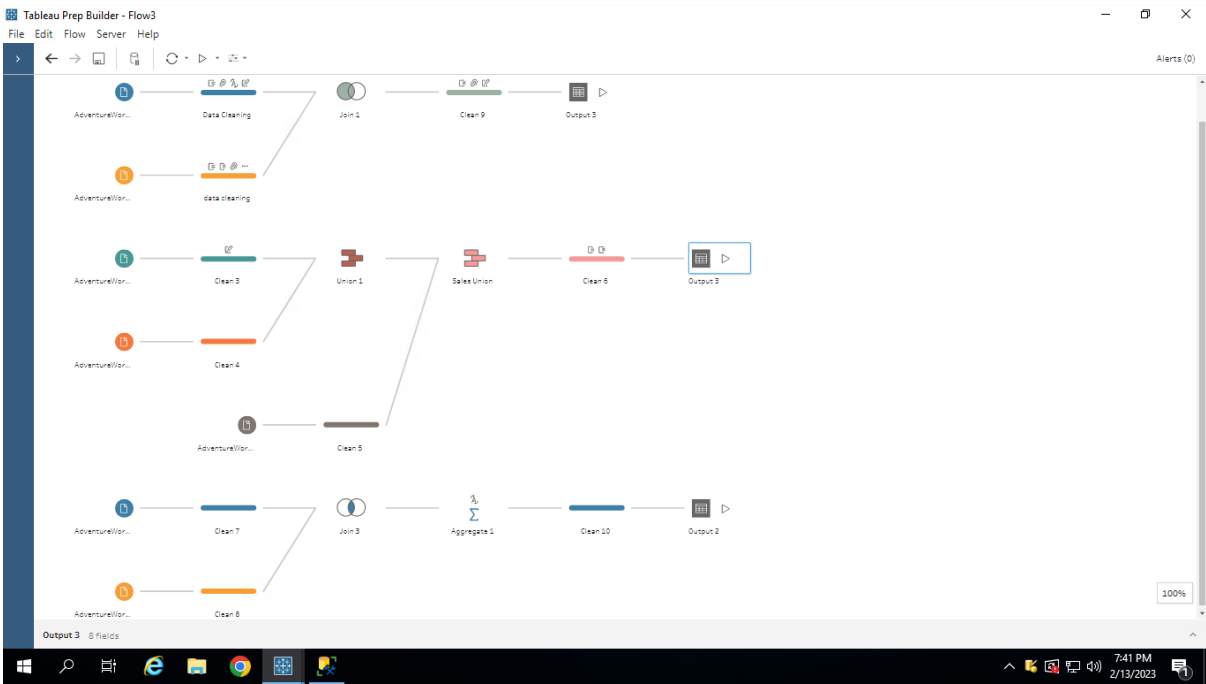


FUNDAMENTALS OF DATA ENGINEERING.

MAHIM DHUNGEL

Overview: ETL Processes in Tableau



Phase 1

1. Connect to Customer file

Connections

AdventureWorks_Cu...
Text file

Search

Tables

AdventureWorks_Cus...

Input

Settings Tables Data Sample Changes (0)

AdventureWorks_Customers 13 fields | Filter Values...

Clear the check box to remove fields. You can also filter your data or change data types. [Add a clean step](#) to view and clean data.

Fields selected: 13 of 13

Field Name	Type	Changes	Preview
CustomerKey	#		11,000, 11,001, 1
Prefix	Abc		MR., Mr.
FirstName	Abc		123JON, EUGENE
LastName	Abc		YANG, HUANG, TC
BirthDate	Abc		null, 05/14/1965,
MaritalStatus	Abc		M, S
Gender	Abc		M
EmailAddress	Abc		jon24@adventure
AnnualIncome	#		90,000, 60,000

2. Add a "Clean Step"

Connections

AdventureWorks_Cu...
Text file

Search

Tables

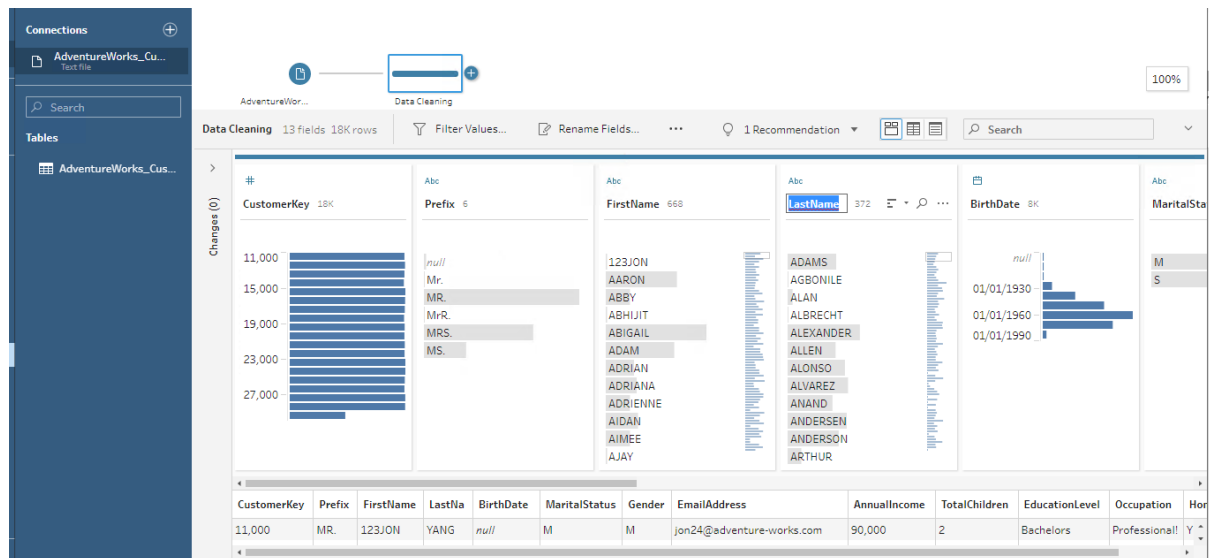
AdventureWorks_Cus...

Data Cleaning 13 fields 18K rows | Filter Values... | Rename Fields... | 1 Recommendation

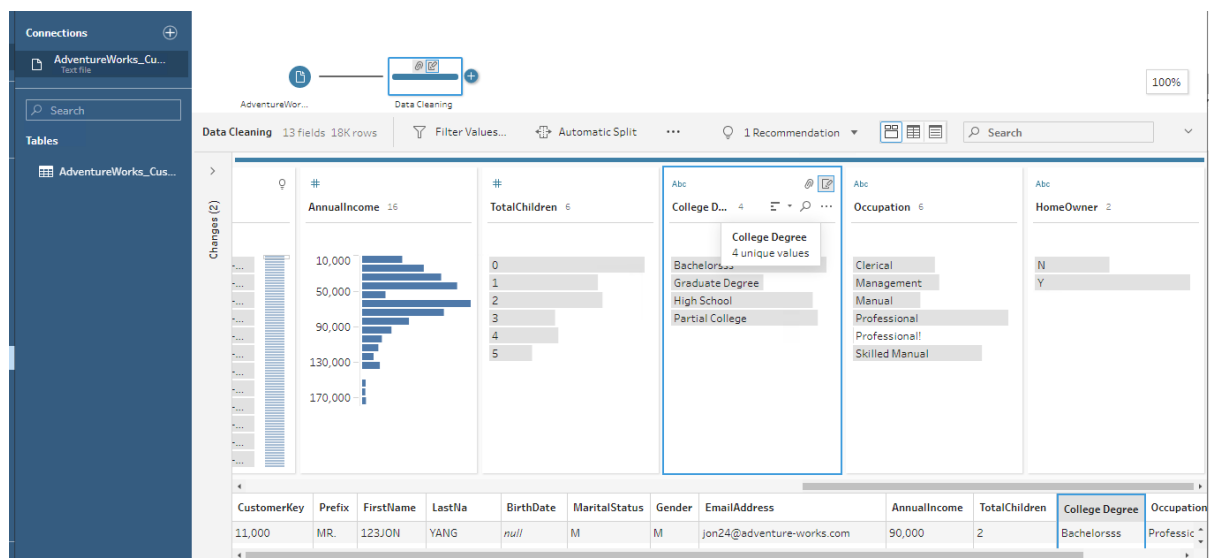
CustomerKey 18K | Prefix 6 | FirstName 668 | LastName 372 | BirthDate 8K | MaritalStatus 6

CustomerKey	Prefix	FirstName	LastName	BirthDate	MaritalStatus	Gender	EmailAddress	AnnualIncome	TotalChildren	EducationLevel	Occupation	HomePhone
11,000	MR.	123JON	YANG	null	M	M	jon24@adventure-works.com	90,000	2	Bachelors	Professional	

3. Rename Field LastNa to LastName



4. Group Education and Replace by College Degree



5. Change marital status M and S to “Married” and “Single”

Data Cleaning

18K rows

✓

Keep Only

✕

Exclude

...

Abc

MaritalSt... 2

MarriedSingle

Abc

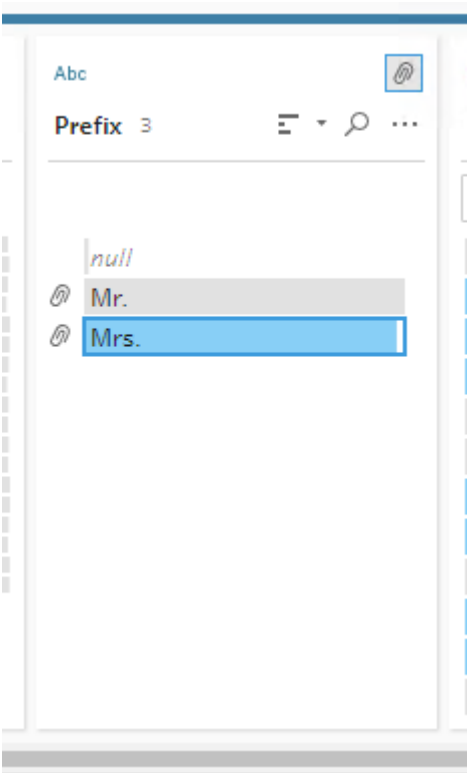
Gender 3

F

M

NA

6. Clean Prefix MrR to MR



7. Remove Numbers from FirstName

Data Cleaning

18K rows | Filter Values... | Automatic Split | 1 Recommendation | Search

Prefix	FirstName	LastName	BirthDate	MaritalStatus	Gender	EmailAddress	AnnualIncome	TotalChildren	College Degree
Mr.	AARON	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ABBY	YANG	1980-01-01	Married	F	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ABHIJIT	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ABIGAIL	YANG	1980-01-01	Married	F	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ADAM	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ADRIAN	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ADRIANA	YANG	1980-01-01	Married	F	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ADRIENNE	YANG	1980-01-01	Married	F	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	AIDAN	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	AIMEE	YANG	1980-01-01	Married	F	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	AJAY	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss
Mrs.	ALAN	YANG	1980-01-01	Married	M	jon24@adventure-works.com	50,000	2	Bachelorss

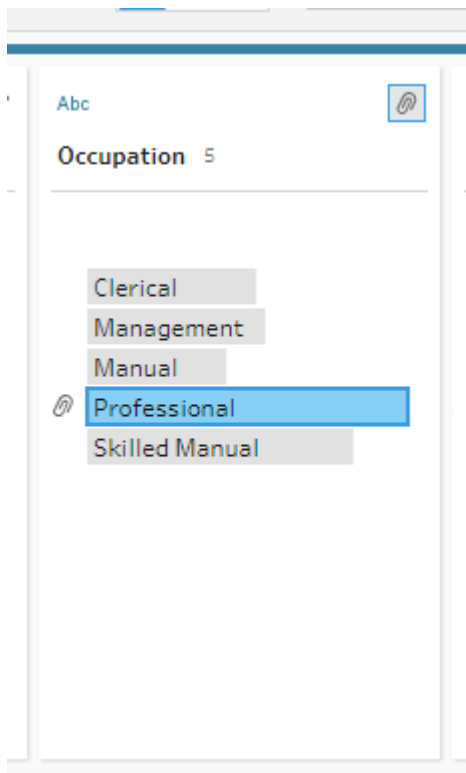
Context Menu (Right-click on FirstName):

- Filter
- Clean
- Group Values
- Split Values
- View State
 - Detail
 - Summary
- Rename Field
- Duplicate Field
- Keep Only Field
- Create Calculated Field
- Publish as Data Role...
- Hide Field
- Remove

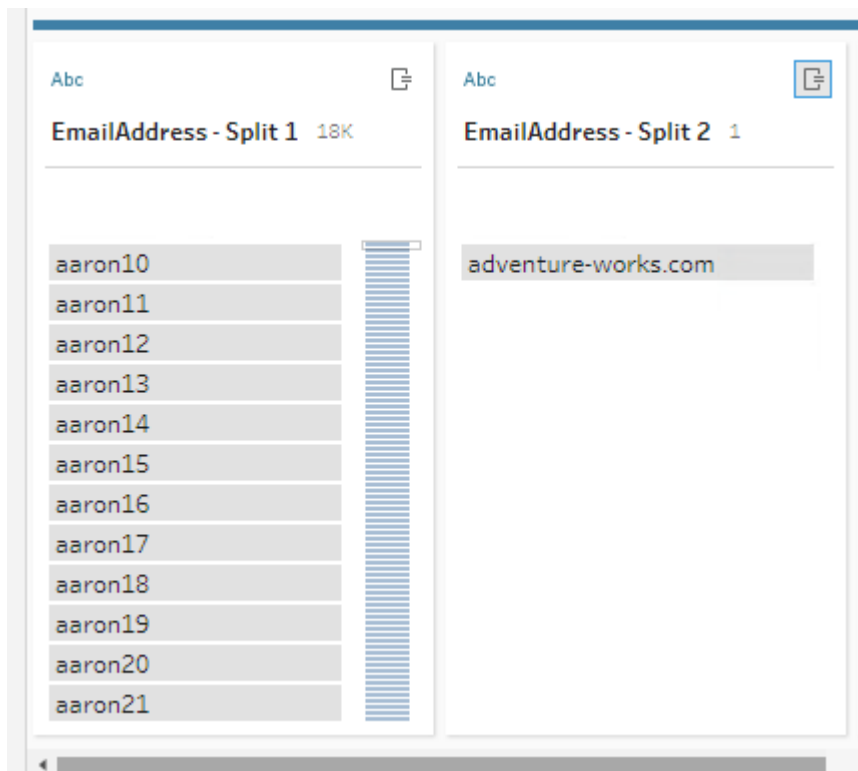
Clean Menu (Right-click on Clean):

- Make Uppercase
- Make Lowercase
- Remove Letters
- Remove Numbers
- Remove Punctuation
- Trim Spaces
- Remove Extra Spaces
- Remove All Spaces

8. Remove punctuations from Occupation

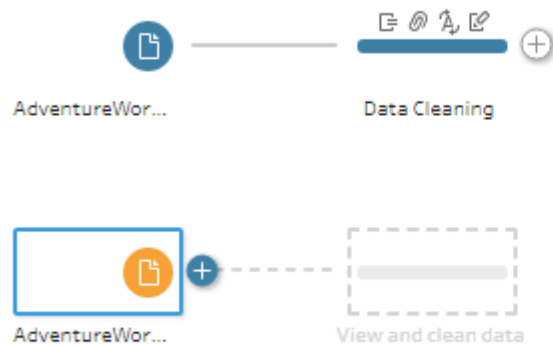


9. Split email address and remove the first letters prior to @



Phase 2

10. Add Adventure Works Customer New file



11. Replace Social Media Account Nulls with NoSocialMedia

The screenshot shows the Power BI Data Cleaning interface. On the left, a list of 'CustomerKey' values (11000 to 11012) is displayed. The main area is titled 'Filter: Null Values' and shows a table with 5 rows. The first row is 'NoSocialMedia', which is highlighted. A tooltip for the 'NoSocialMedia' row shows 'null' and '16,074 (89%) rows'. The second row is 'FB', the third is 'FB, Instagram', the fourth is 'Instagram', and the fifth is 'Twitter'. A tooltip for the 'Instagram' row shows '16,074 (100%) highlighted'. On the right, the 'Keep Only' section is visible, with options to 'Select values to keep.' and radio buttons for 'Null values' and 'Non-null values'.

CustomerKey	Social Media Accounts
11000	NoSocialMedia
11001	FB
11002	FB, Instagram
11003	Instagram
11004	Twitter
11005	
11007	
11008	
11009	
11010	
11011	
11012	

12. Split Social Media Fields

The screenshot shows a data visualization interface with a split view. The left pane, titled 'Social Me...', displays a list of social media categories: FB, Instagram, NoSocialMedia, and Twitter. The right pane, titled 'CustomerKey', displays a list of values from 11000 to 11012. The top toolbar includes options like 'Filter Values...', 'Rename Fields...', and a search bar. The interface is designed for data exploration and analysis.

13. Load Data toDimCustomer Tables

	Social Media Accounts Information	CustomerKey
1	FB	11000
2	FB	11001
3	FB	11002
4	FB	11003
5	FB	11004
6	FB	11005
7	FB	11007
8	FB	11008
9	FB	11009
10	FB	11010
11	FB	11011

Query executed successfully. RAS-RDSH-04\SQLEXPRESS (15.... CLARKU\MDhungel (57)

Phase 3

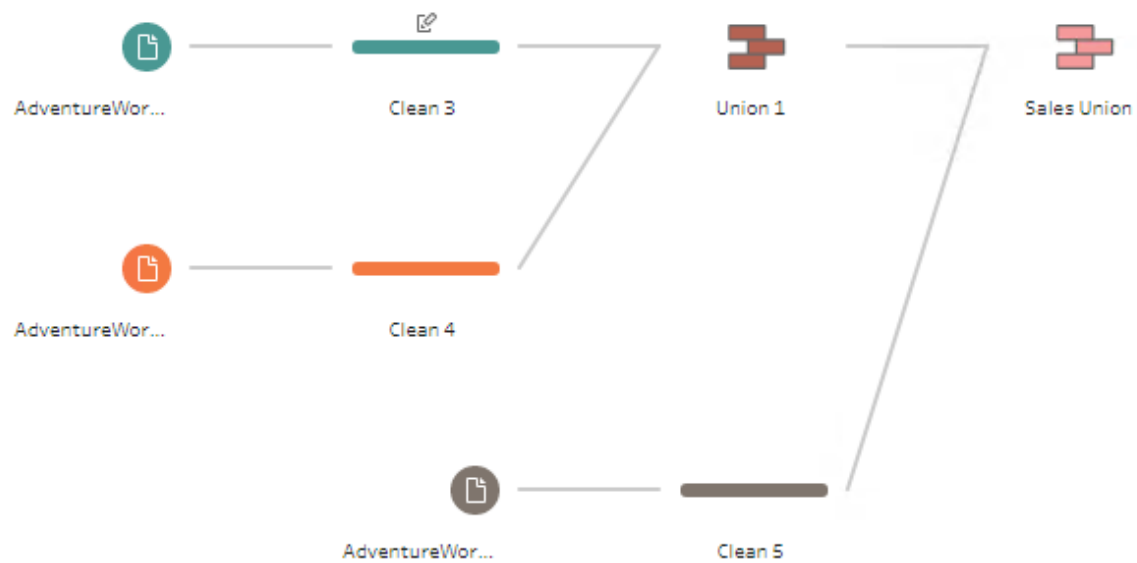
- Add AdventureWorks_Sales_2015
- Add AdventureWorks_Sales_2016
- Add AdventureWorks_Sales_2017



14. Clean OrderQuantity from "Quantite de Ventes" to "OrderQuantity"

A screenshot of a data table. The table has a header row with a blue hash symbol (#) and a blue document icon. The first column is labeled "OrderQuantity" and the second column is labeled "1". The table contains a single row with the value "1" in the "OrderQuantity" column. The table is titled "OrderQuantity" at the bottom.

15. Union all three Files



16. Add a calculation to extract "Year" from "Order Date"

OrderDate 911 StockDate 1K OrderNumber 25K

06/19/2017
06/20/2017
06/21/2017
06/22/2017
06/23/2017
06/24/2017
06/25/2017
06/26/2017
06/27/2017
06/28/2017
06/29/2017
06/30/2017

SO45070

SO45070

SO45070

customerKey TerritoryKey OrderLineIt

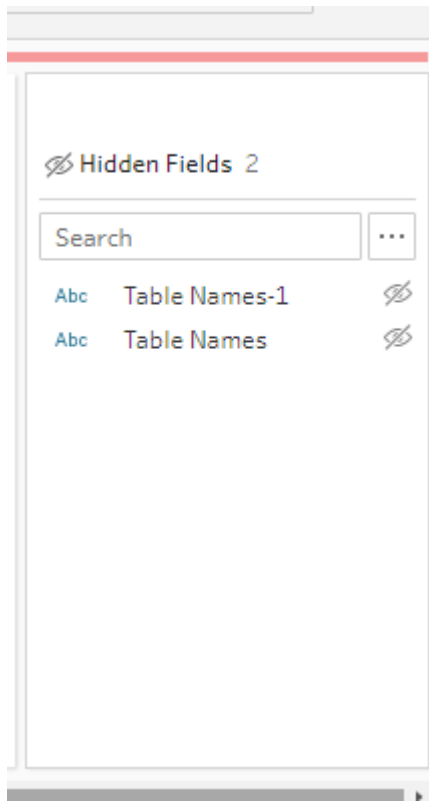
01/01/2017	12/13/2003	SO61285	214	23,791	1	2
01/01/2017	09/24/2003	SO61285	214	23,791	1	3

Filter
Clean
Group Values
Convert Dates
Split Values
View State
✓ Detail
Summary
Rename Field
Duplicate Field
Keep Only Field
Create Calculated Field
Publish as Data Role...
Hide Field
Remove

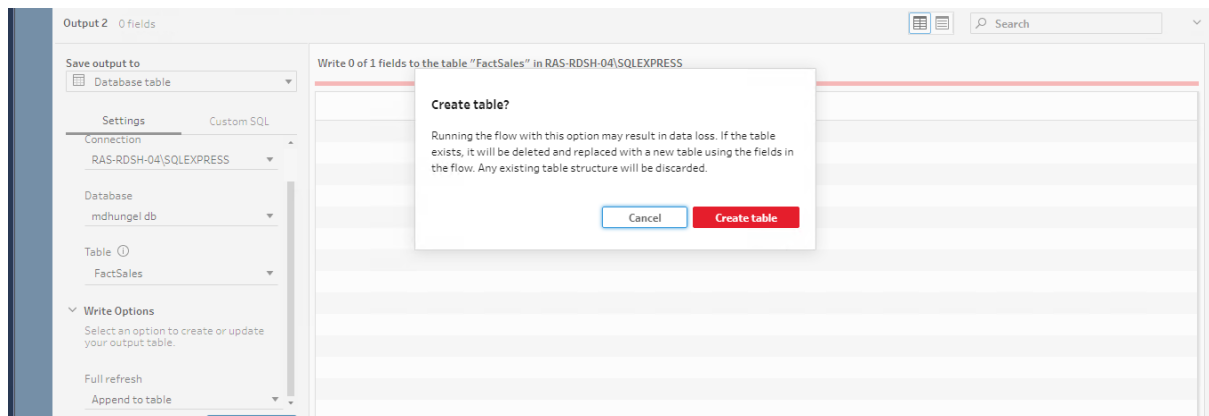
Date and Time
Year Number
Quarter Number
Month Number
Month Name
Week Number
Day of the Week
Day of the Month
Custom Fiscal Year

Examples:
2015
2016
2017

17. Remove Table Names

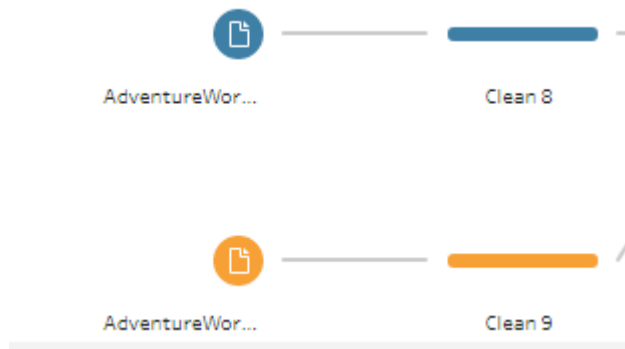


18. Load Data into FactSales table

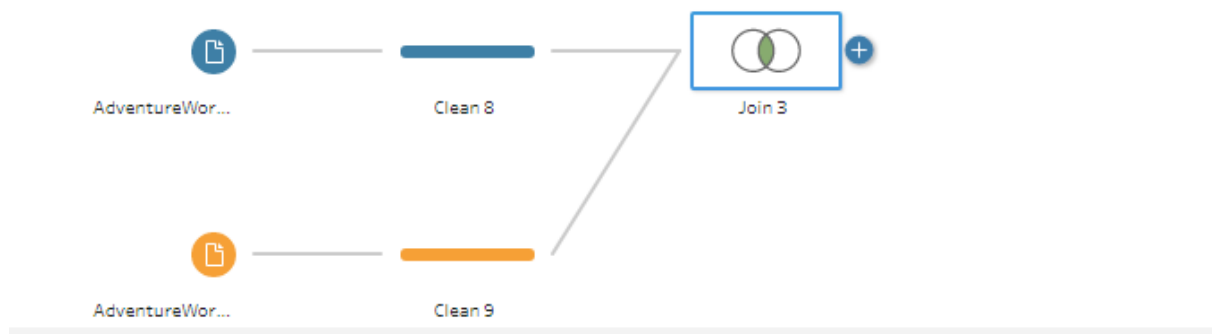


Phase 4

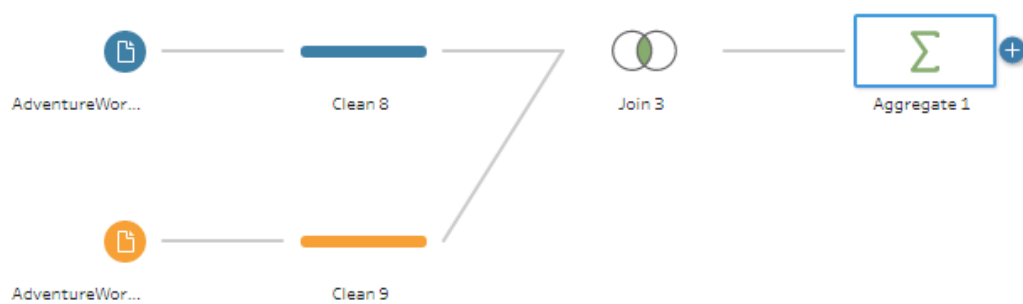
19. Add Returns table and Add Product Table



20. Inner Join on Product key



21. Add aggregation



22. Sum of Prices for Product Returns

Abc	GROUP	ModelName
Abc	GROUP	ProductColor
#	SUM	ProductCost
Abc	GROUP	ProductDescription
#	SUM	ProductKey
#	SUM	ProductKey-1
Abc	GROUP	ProductName
#	SUM	ProductPrice
Abc	GROUP	ProductSize

23. Create a new table with the following columns:

- Product Key
- Territory Key

Aggregate 12 fields 518 rows

Filter Values...

Settings

Changes (2)

Additional Fields

Drag fields to aggregate or group them.

Search

Add All

Remove All

Abc

GROUP

ModelName

Abc

GROUP

ProductColor

#

SUM

ProductCost

Abc

GROUP

ProductDescription

#

SUM

ProductKey-1

Abc

GROUP

ProductName

#

SUM

ProductPrice

Abc

GROUP

ProductSize

Abc

GROUP

ProductSKU

Abc

GROUP

ProductStyle

#

SUM

ProductSubcategoryKey

Abc

GROUP

ReturnDate

#

SUM

ReturnQuantity

#

SUM

Number of Rows (Aggregated)

Grouped Fields

Abc

GROUP

TerritoryKey 8

Abc

GROUP

ProductKey 124

1

10

4

5

6

7

8

9

214

215

220

223

226

229

232

235

310

311

312

313

220

7 (1%) rows

TerritoryKey

ProductKey

Aggregated Fields

Drop fields here to aggregate them

24. Create a new table in the database called DMReturns

The screenshot displays the SQL Server Enterprise Manager interface. On the left, the 'Database Diagrams' and 'Tables' folders are expanded. Under 'Tables', the 'dbo.DMReturns' table is highlighted. The right pane shows the 'Results' tab with a query result set. The query result set has four columns: 'TerritoryKey', 'ProductKey', and 'ProductCost'. The data is as follows:

	TerritoryKey	ProductKey	ProductCost
1	7	377	2641.3676
2	8	350	1898.0944
3	8	581	2165.02
4	9	592	308.2179
5	4	575	2963.8758
6	4	482	10.0869
7	10	576	1481.9379
8	1	475	26.1763
9	9	215	72.1668
10	4	214	209.3808
11	10	563	1481.9379

At the bottom of the interface, a status bar indicates 'Query executed successfully.' and the status is 'Ready'.