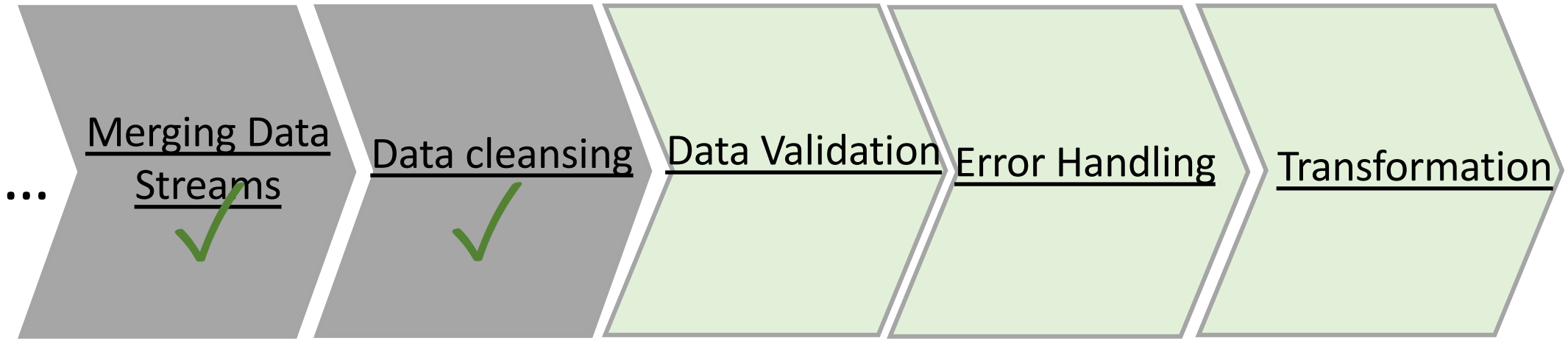
DATA ANALYTICS
(Data Warehouse)
Pentaho Data Integration

Luca Cinelli, PhD

luca.cinelli@unical.it

Outline


...
Merging Data Streams
Data cleansing
Data Validation
Error Handling
Transformation

Data Validation

Data Validation

- **Data Cleansing** is to correct mistakes and format for data to improve its quality.
- **Data Validation** is to ensure that the **Data** complies with the business **rules**
- **Examples**:
  - Age field should contain integer type values
  - Customer age should be more than 18 years
  - Product ID in the sales table should be available in the Product table as well
  - Credit Card number/ email ID/ Phone number should have a predefined format

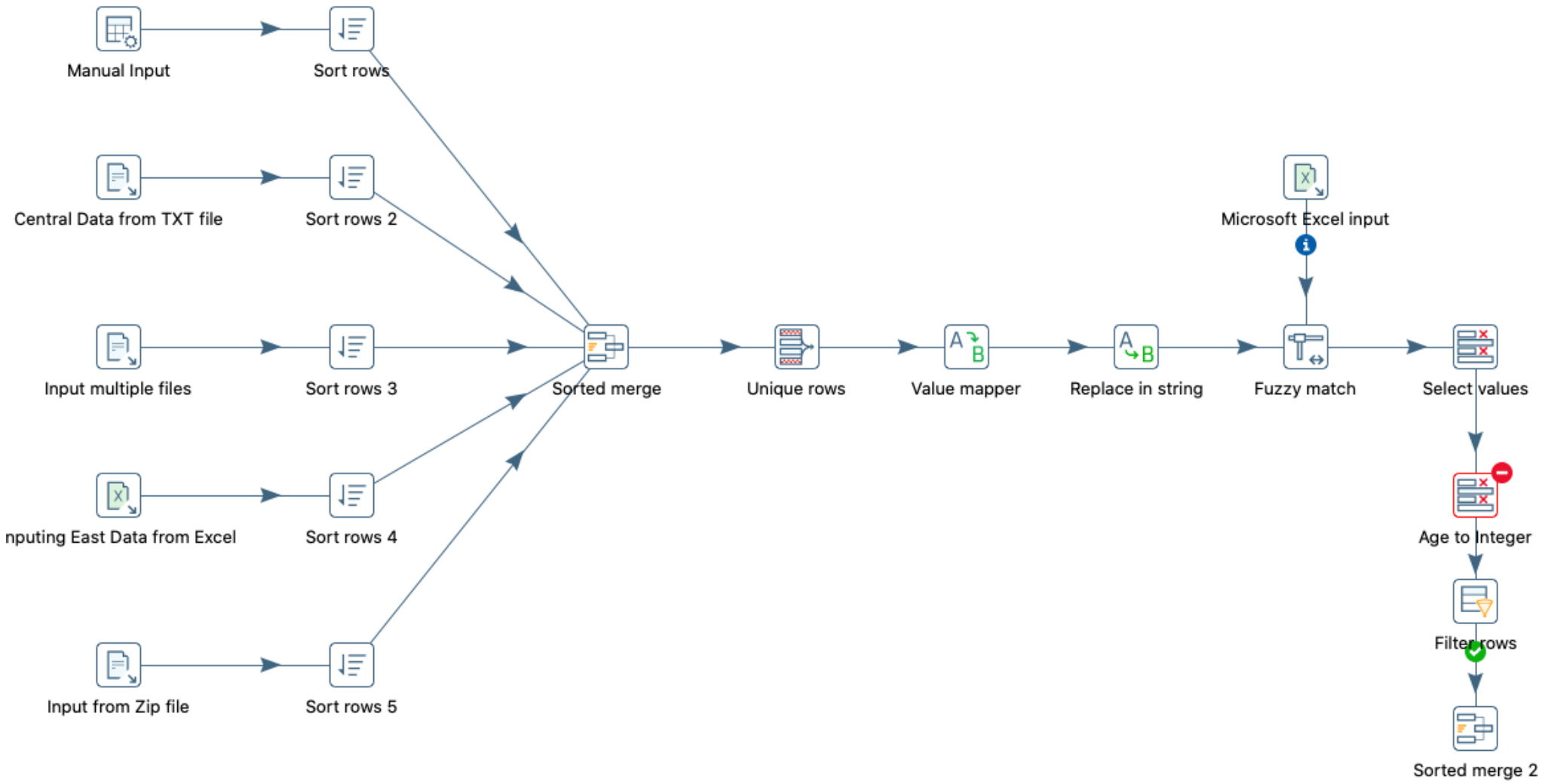
Data Validation: Customer data example

Age imported as String format should be integer and >=0

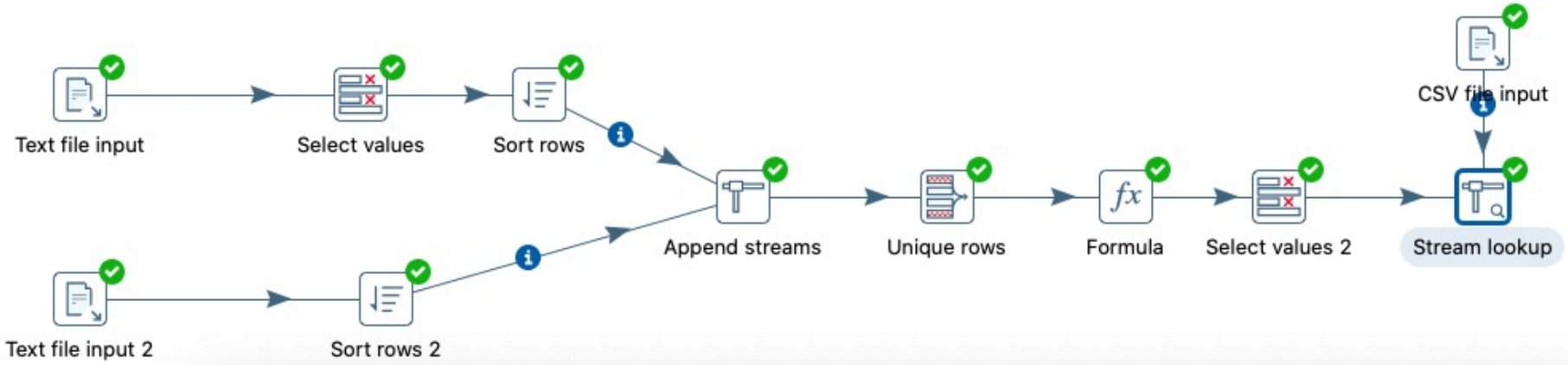
CustomerComplete

	Customer ID	Customer Name	Segment	Age	Country	City	State	Postal Code	Region
2	SO-20335	Sean O'Donnell	Consumer	65	United States	Fort Lauderdale	Florida	33311	South
3	BH-11710	Brosina Hoffman	Consumer	20	United States	Los Angeles	California	90032	West
4	AA-10480	Andrew Allen	Consumer	50	United States	Concord	North Carolina	28027	South
5	IM-15070	Irene Maddox	Consumer	66	United States	Seattle	Washington	98103	West
6	HP-14815	Harold Pawlan	Home Office	20	United States	Fort Worth	Texas	76106	Central
7	PK-19075	Pete Kriz	Consumer	46	United States	Madison	Wisconsin	53711	Central
8	AG-10270	Alejandro Grove	Consumer	18	United States	West Jordan	Utah	84084	West
9	ZD-21925	Zuschuss Donatelli	Consumer	66	United States	San Francisco	California	94109	West
10	KB-16585	Ken Black	Corporate	67	United States	Fremont	Nebraska	68025	Central
11	SF-20065	Sandra Flanagan	Consumer	41	United States	Philadelphia	Pennsylvania	19140	East
12	EB-13870	Emily Burns	Consumer	34	United States	Orem	Utah	84057	West
13	EH-13945	Eric Hoffmann	Consumer	21	United States	Los Angeles	California	90049	West
14	TB-21520	Tracy Blumstein	Consumer	48	United States	Philadelphia	Pennsylvania	19140	East
15	MA-17560	Matt Abelman	Home Office	19	United States	Houston	Texas	77095	Central
16	GH-14485	Gene Hale	Corporate	28	United States	Richardson	Texas	75080	Central
17	SN-20710	Steve Nguyen	Home Office	46	United States	Houston	Texas	77041	Central
18	LC-16930	Linda Cazamias	Corporate	31	United States	Naperville	Illinois	60540	Central
19	RA-19885	Ruben Ausman	Corporate	51	United States	Los Angeles	Cakifornia	90049	West
20	ES-14080	Erin Smith	Corporate	2o	United States	Melbourne	Florida	32935	South
21	ON-18715	Odella Nelson	Corporate	27	United States	Eagan	Minnesota	55122	Central
22	PO-18865	Patrick O'Donnell	Consumer	64	United States	Westland	Michigan	48185	Central
23	LH-16900	Lena Hernandez	Consumer	66	United States	#Dover	Delaware	19901	East

Data Validation: Customer data example



Data Validation: example - Sales refer to products

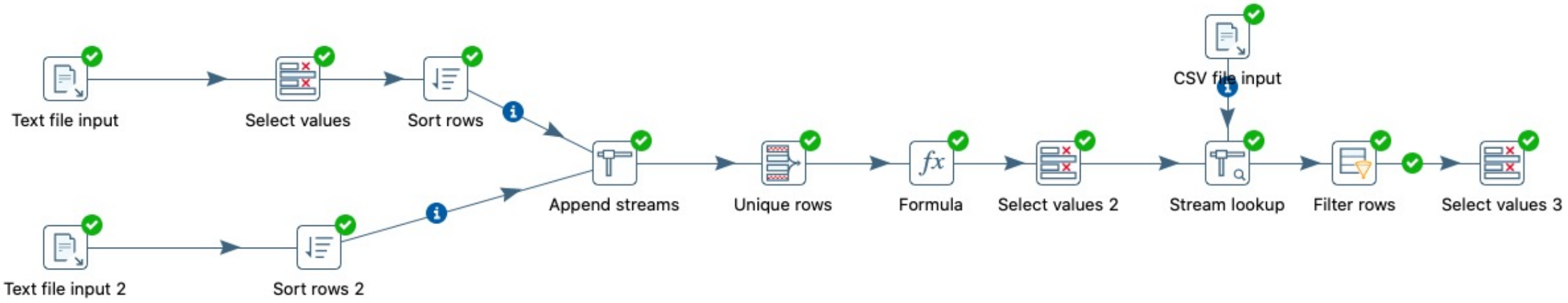


Examine preview data

Rows of step: Stream lookup (9994 rows)

#	Order_Line	Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Product_ID	Sales	Quantity	Discount	Profit	Product Name
8...	8299	US-2017-169...	14-02-20...	18-02-2...	Standard Class	MG-17650	OFF-SU-10004...	21.8	3	0	5.9	Acme Stainless Steel Office Ships
8...	8300	US-2017-169...	14-02-20...	18-02-2...	Standard Class	MG-17650	OFF-AP-100019...	91.6	5	0	26.6	Acco 6 Outlet Guardian Premium Plu
8...	8301	CA-2017-144...	15-02-20...	18-02-2...	First Class	CP-12085	OFF-LA-10004544	47.4	4	20	17.8	Avery 505
8...	8302	CA-2017-144...	15-02-20...	18-02-2...	First Class	CP-12085	OFF-ST-10004507	27.4	2	20	2.4	Advantus Rolling Storage Box
8...	8303	CA-2017-144...	15-02-20...	18-02-2...	First Class	CP-12085	OFF-BI-10002012	3.2	9	80	-5.2	Wilson Jones Easy Flow II Sheet Lifte
8...	8304	CA-2017-104...	15-02-20...	20-02-2...	Second Class	MD-17860	OFF-AR-11001...	9.4	7	20	0.7	<null>
8...	8305	CA-2017-101...	15-02-20...	21-02-2...	Standard Class	RW-19540	FUR-FU-10003...	148	3	0	41.5	Electrix Fluorescent Magnifier Lamps
8...	8306	CA-2017-159...	17-02-20...	22-02-2...	Standard Class	JF-15490	OFF-ST-10000344	10.7	1	20	0.8	Neat Ideas Personal Hanging Folder

Data Validation: example - Sales refer to products



Execution Results

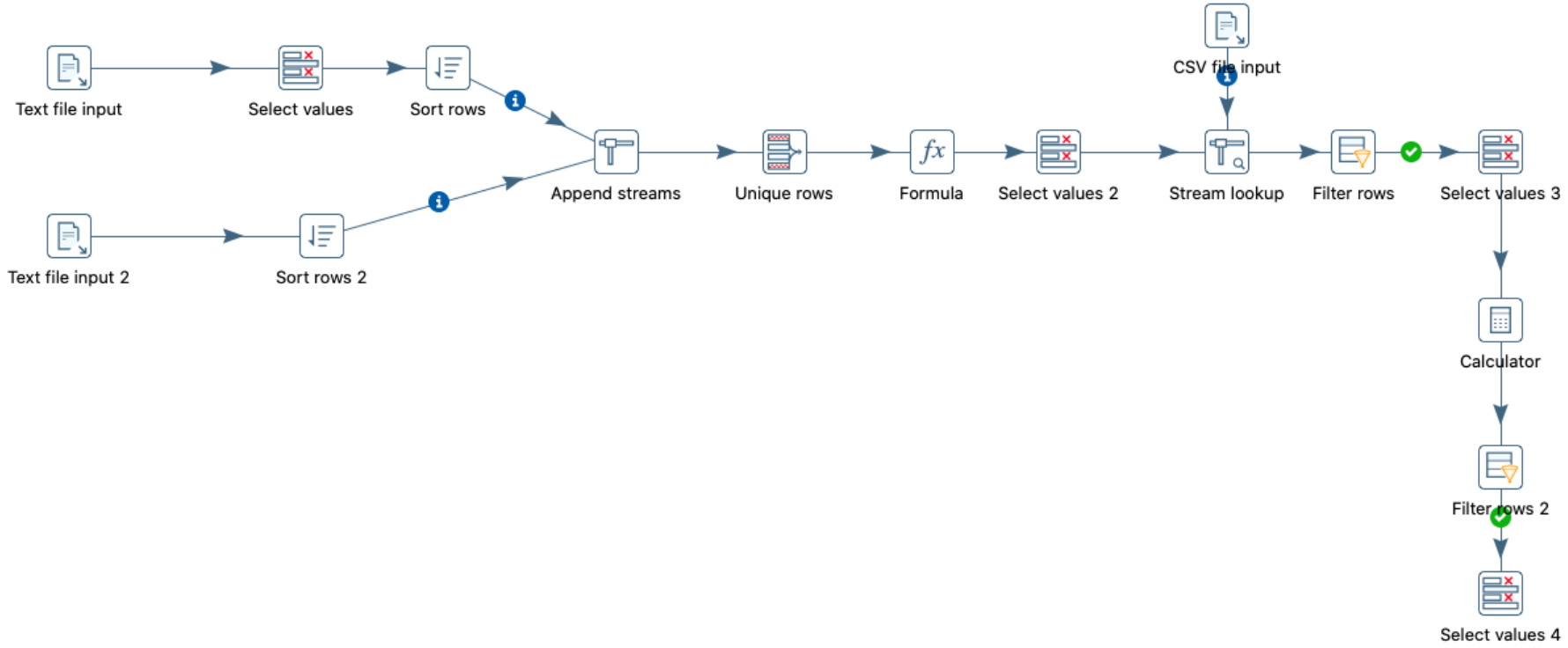
Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	Order_Line	Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Product_ID	Sales	Quantity	Discount	Profit
1	1	CA-2014-103...	21-06-20...	25-06-2...	Standard Class	DP-13000	OFF-PA-100001...	16.4	2	20	5.5
2	2	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-LA-10003223	11.8	3	20	4.3
3	3	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-ST-10002743	272.7	3	20	-64.8
4	4	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-BI-10004094	3.5	2	80	-5.5
5	5	CA-2014-141...	23-06-20...	30-06-2...	Standard Class	MB-18085	OFF-AR-10003...	19.5	3	20	4.9
6	6	CA-2014-130...	24-06-20...	26-06-2...	Second Class	LS-17230	OFF-PA-100020...	19.4	3	0	9.3

Data Validation: examples on dates – sales data

- the order day >= the ship date values



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

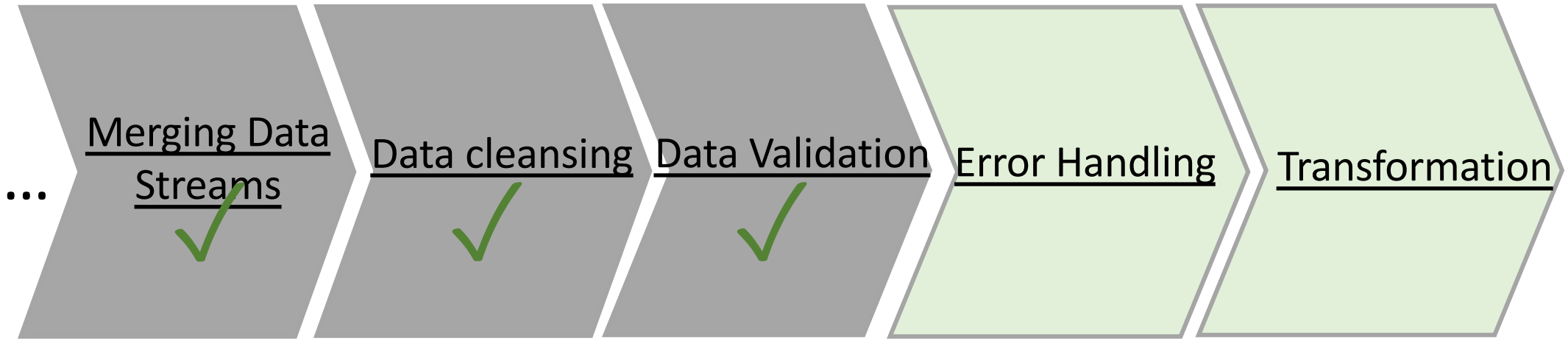
First rows | Last rows | Off

#	Order_Line	Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Product_ID	Sales	Quantity	Discount	Profit	number of days
1	1	CA-2014-103...	21-06-20...	25-06-2...	Standard Class	DP-13000	OFF-PA-100001...	16.4	2	20	5.5	4
2	2	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-LA-10003223	11.8	3	20	4.3	4
3	3	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-ST-10002743	272.7	3	20	-64.8	4
4	4	CA-2014-112...	22-06-20...	26-06-2...	Standard Class	PO-19195	OFF-BI-10004094	3.5	2	80	-5.5	4
5	5	CA-2014-141...	23-06-20...	30-06-2...	Standard Class	MB-18085	OFF-AR-10003...	19.5	3	20	4.9	7
6	6	CA-2014-130	24-06-20	26-06-2	Second Class	LS-17230	OFF-PA-100020	19.4	3	0	9.3	2

Data Validation: Common steps

Scenario	Step
Value must have a given **data type** such as String or Date	Select values
Value cannot be **null**	Filter rows
Numbers or dates should fall inside an expected **range**	Filter rows (>, <, or = functions)
Values must belong to **list** found in an external source such as a file or a database	Stream Lookup or Data base Lookup
Text should not contain certain terms or substrings	Replace in string step

Outline


...
Merging Data Streams
Data cleansing
Data Validation
Error Handling
Transformation

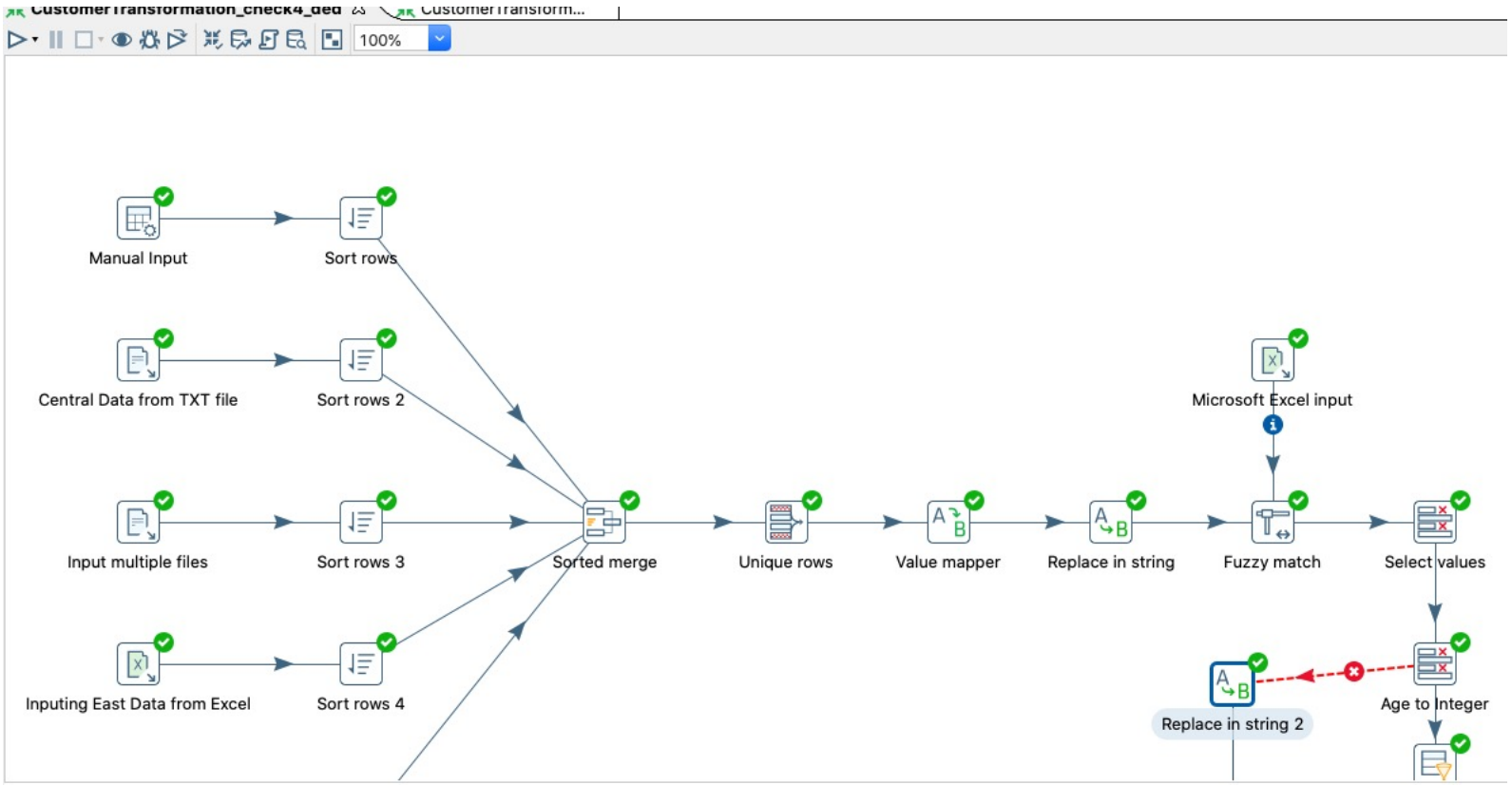
Error Handling

Data Validation: Error Handling

- If some rows do not respect the data validation rules, those are the error rows.
- We need to properly handle error rows.
- Error can be handled in these four ways:
  1. Discarding the error rows
  2. Separating error rows, processing them and remerging them with the main stream
  3. Reporting the error rows to the log
  4. Writing the error rows in a file or a dedicated table for further revision

Data Validation: Error Handling – customer data example

Separating error rows, processing them, and remerging them with the main stream



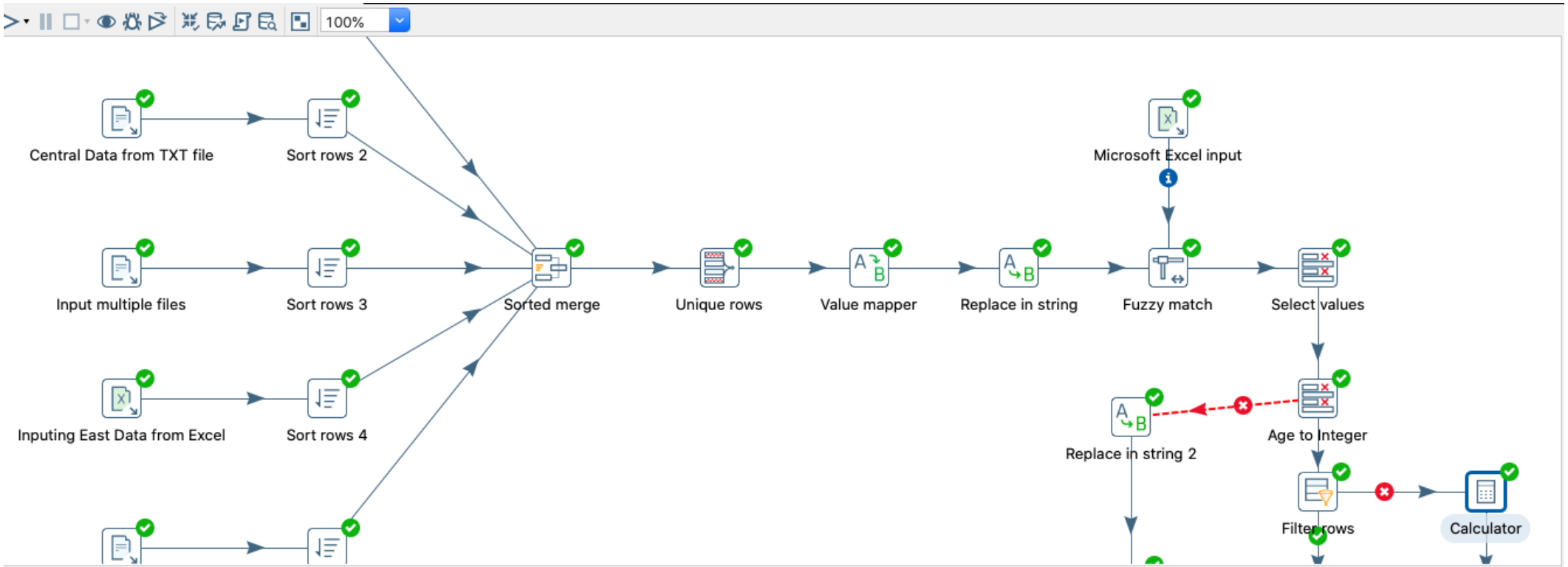
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	Customer ID	Customer Name	Segment	Age	City	State	Country	Postal Code	Region
1	CC-12100	Chad Cunningham	Home Office	30	Columbus	Indiana	United States	47201	Central
2	ES-14080	Erin Smith	Corporate	20	Melbourne	Florida	United States	32935	South
3	JE-15475	Jeremy Ellison	Consumer	30	Troy	New York	United States	12180	East
4	NG-18430	Nathan Gelder	Consumer	50	Cincinnati	Ohio	United States	45231	East

Data Validation: Error Handling – customer data example



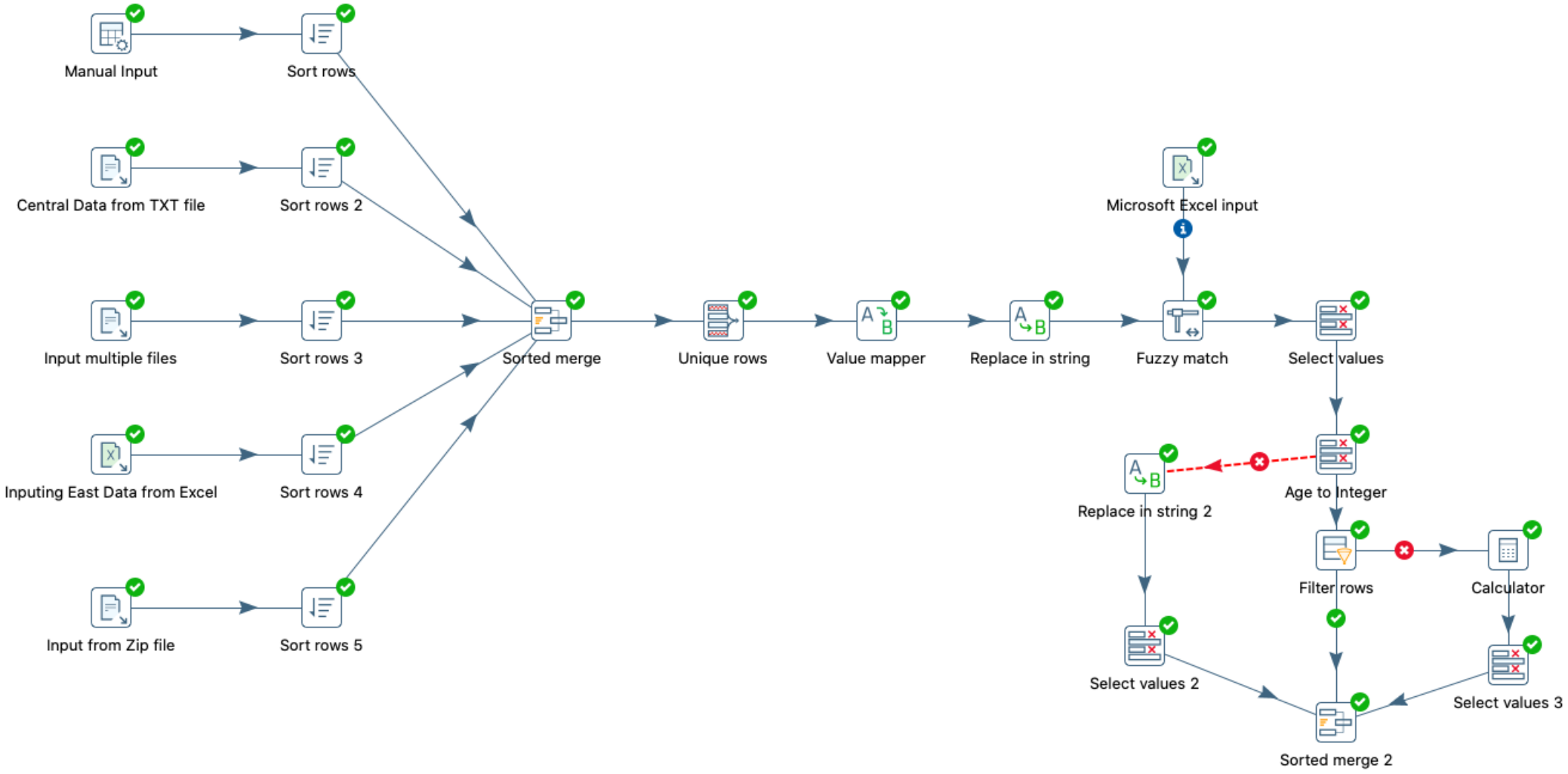
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

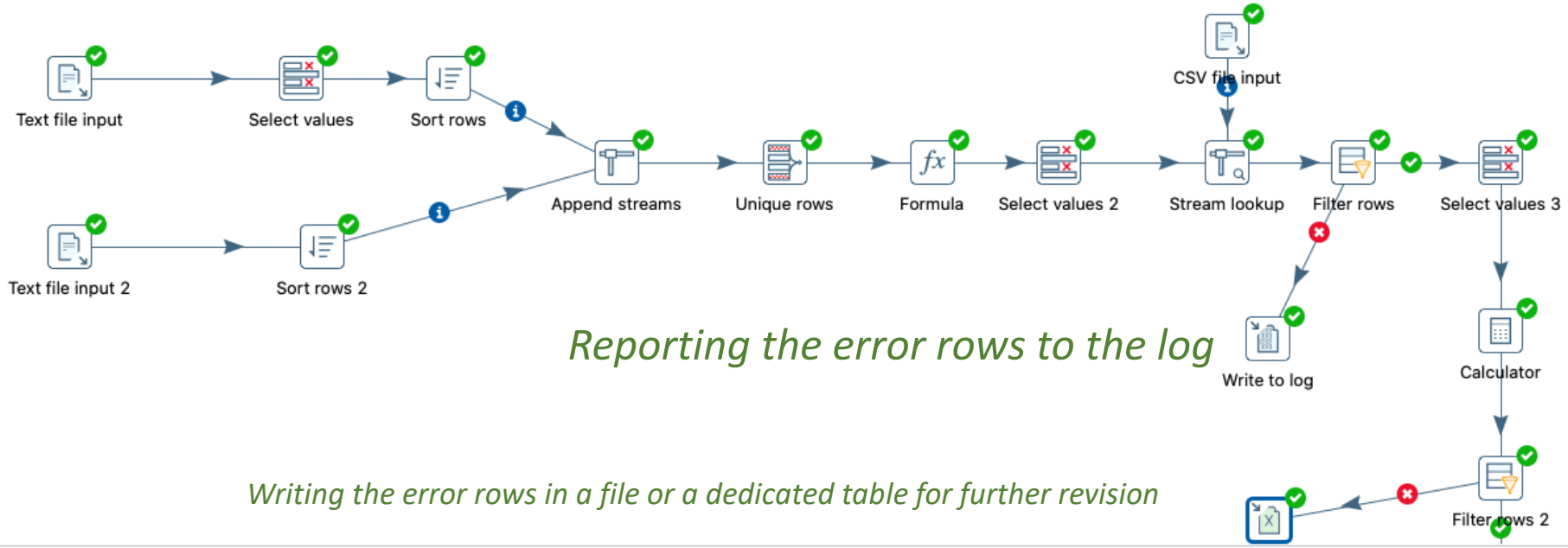
First rows | Last rows | Off

#	Customer ID	Customer Name	Segment	Age	City	State	Country	Postal Code	Region	Age1
1	AG-10765	Anthony Garverick	Home Office	-40	Philadelphia	Pennsylvania	United States	19120	East	40
2	EG-13900	Emily Grady	Consumer	-60	Oceanside	New York	United States	11572	East	60
3	MC-17425	Mark Cousins	Corporate	-50	Mobile	Alabama	United States	36608	South	50

Data Validation: Error Handling – customer data example



Data Validation: Error Handling – example - Sales refer to products



Reporting the error rows to the log

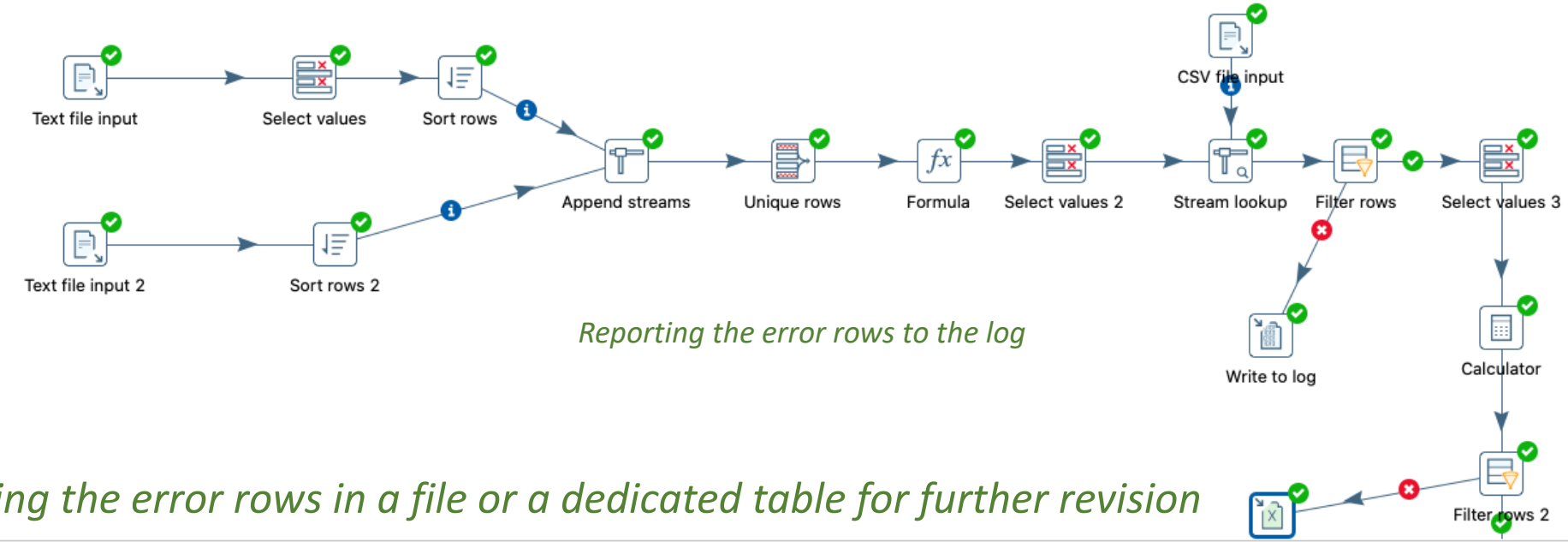
Writing the error rows in a file or a dedicated table for further revision

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

```
2020/11/14 22:06:52 - Write to log.0 - ------------> Linenr 1------------------------------
2020/11/14 22:06:52 - Write to log.0 - Refernce of product ID not found in Product Lookup
2020/11/14 22:06:52 - Write to log.0 -
2020/11/14 22:06:52 - Write to log.0 - Order_Line = 8304
2020/11/14 22:06:52 - Write to log.0 - Order_ID = CA-2017-104024
2020/11/14 22:06:52 - Write to log.0 - Order_Date = 15-02-2020
2020/11/14 22:06:52 - Write to log.0 - Ship_Date = 20-02-2020
2020/11/14 22:06:52 - Write to log.0 - Ship_Mode = Second Class
2020/11/14 22:06:52 - Write to log.0 - Customer_ID = MD-17860
2020/11/14 22:06:52 - Write to log.0 - Product_ID = OFF-AR-11001972
2020/11/14 22:06:52 - Write to log.0 - Sales = 9.4
2020/11/14 22:06:52 - Write to log.0 - Quantity = 7
2020/11/14 22:06:52 - Write to log.0 - Discount = 20
2020/11/14 22:06:52 - Write to log.0 - Profit = 0.7
```

Data Validation: Error Handling – examples on dates – sales data



Reporting the error rows to the log

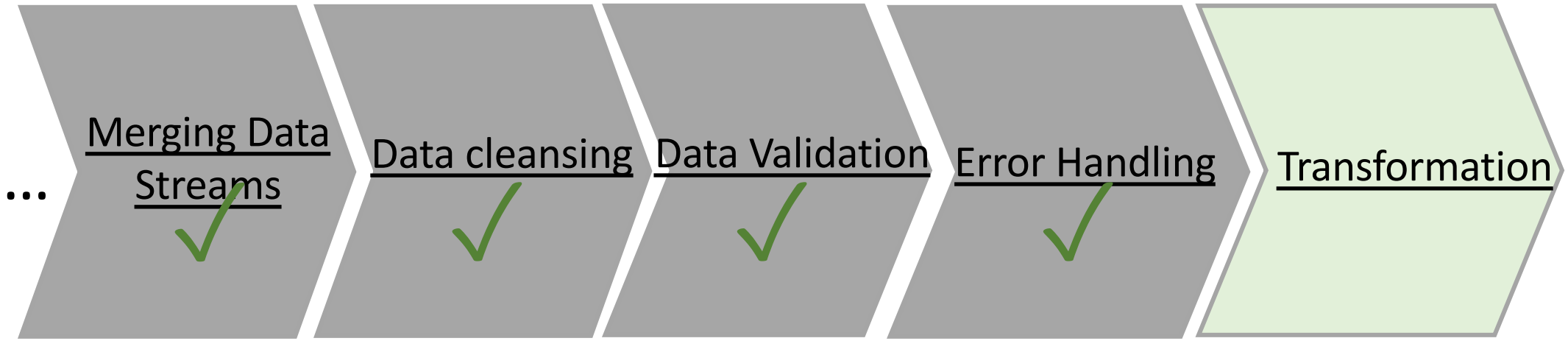
Writing the error rows in a file or a dedicated table for further revision

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

```
2020/11/14 22:06:52 - Write to log.0 - ------------> Linenr 1------------------------------
2020/11/14 22:06:52 - Write to log.0 - Refernce of product ID not found in Product Lookup
2020/11/14 22:06:52 - Write to log.0 -
2020/11/14 22:06:52 - Write to log.0 - Order_Line = 8304
2020/11/14 22:06:52 - Write to log.0 - Order_ID = CA-2017-104024
2020/11/14 22:06:52 - Write to log.0 - Order_Date = 15-02-2020
2020/11/14 22:06:52 - Write to log.0 - Ship_Date = 20-02-2020
2020/11/14 22:06:52 - Write to log.0 - Ship_Mode = Second Class
2020/11/14 22:06:52 - Write to log.0 - Customer_ID = MD-17860
2020/11/14 22:06:52 - Write to log.0 - Product_ID = OFF-AR-11001972
2020/11/14 22:06:52 - Write to log.0 - Sales = 9.4
2020/11/14 22:06:52 - Write to log.0 - Quantity = 7
2020/11/14 22:06:52 - Write to log.0 - Discount = 20
2020/11/14 22:06:52 - Write to log.0 - Profit = 0.7
```

Outline


...
Merging Data Streams
Data cleansing
Data Validation
Error Handling
Transformation

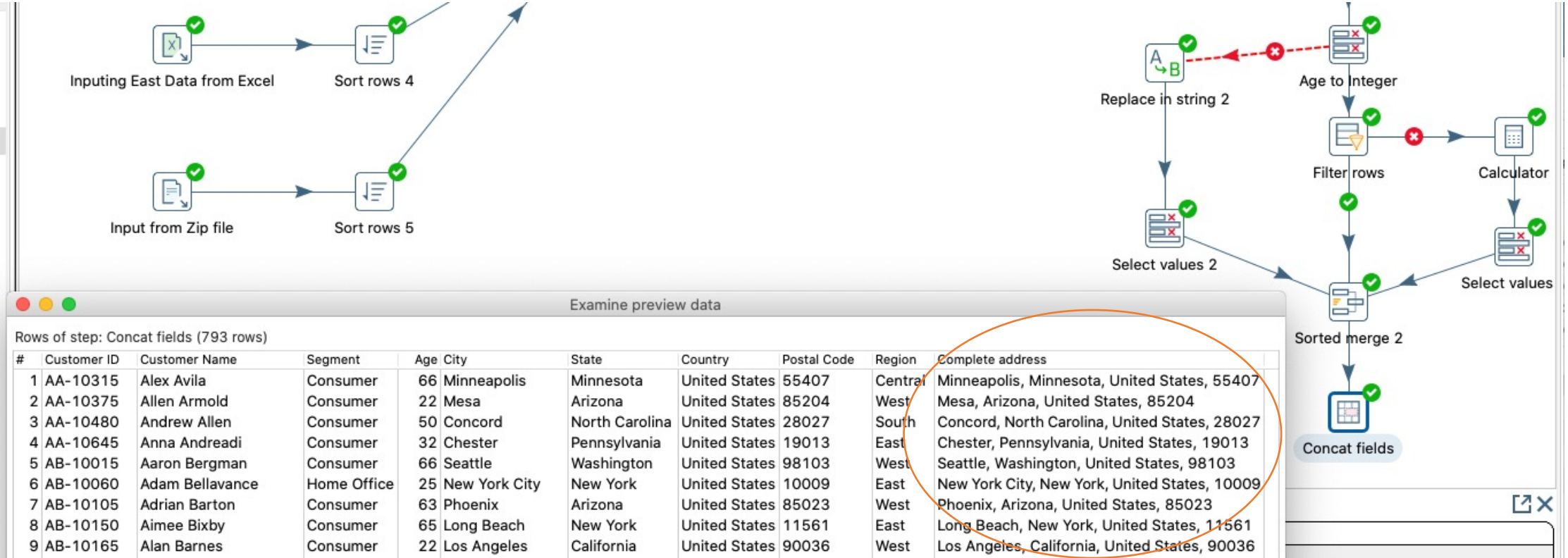
Transformation

Transformation: examples

- Concatenating data
- Data aggregation – group by
- Normalization and denormalization
- Create number ranges

Transformation: Customer data example - Concatenating address field

- Goal: to combine the data contained in several address fields in one field address, such as city, state, country and postal code.
  (CustomerTransformation_check5_ded clean_fuzzy_Validation_errorh_Tconcatenate.ktr)



#	Customer ID	Customer Name	Segment	Age	City	State	Country	Postal Code	Region	Complete address
1	AA-10315	Alex Avila	Consumer	66	Minneapolis	Minnesota	United States	55407	Central	Minneapolis, Minnesota, United States, 55407
2	AA-10375	Allen Armold	Consumer	22	Mesa	Arizona	United States	85204	West	Mesa, Arizona, United States, 85204
3	AA-10480	Andrew Allen	Consumer	50	Concord	North Carolina	United States	28027	South	Concord, North Carolina, United States, 28027
4	AA-10645	Anna Andreadi	Consumer	32	Chester	Pennsylvania	United States	19013	East	Chester, Pennsylvania, United States, 19013
5	AB-10015	Aaron Bergman	Consumer	66	Seattle	Washington	United States	98103	West	Seattle, Washington, United States, 98103
6	AB-10060	Adam Bellavance	Home Office	25	New York City	New York	United States	10009	East	New York City, New York, United States, 10009
7	AB-10105	Adrian Barton	Consumer	63	Phoenix	Arizona	United States	85023	West	Phoenix, Arizona, United States, 85023
8	AB-10150	Aimee Bixby	Consumer	65	Long Beach	New York	United States	11561	East	Long Beach, New York, United States, 11561
9	AB-10165	Alan Barnes	Consumer	22	Los Angeles	California	United States	90036	West	Los Angeles, California, United States, 90036

Transformation: examples

- Concatenating data
- Data aggregation – group by
- Normalization and denormalization
- Create number ranges

Transformation: Data aggregation - group by

Examples

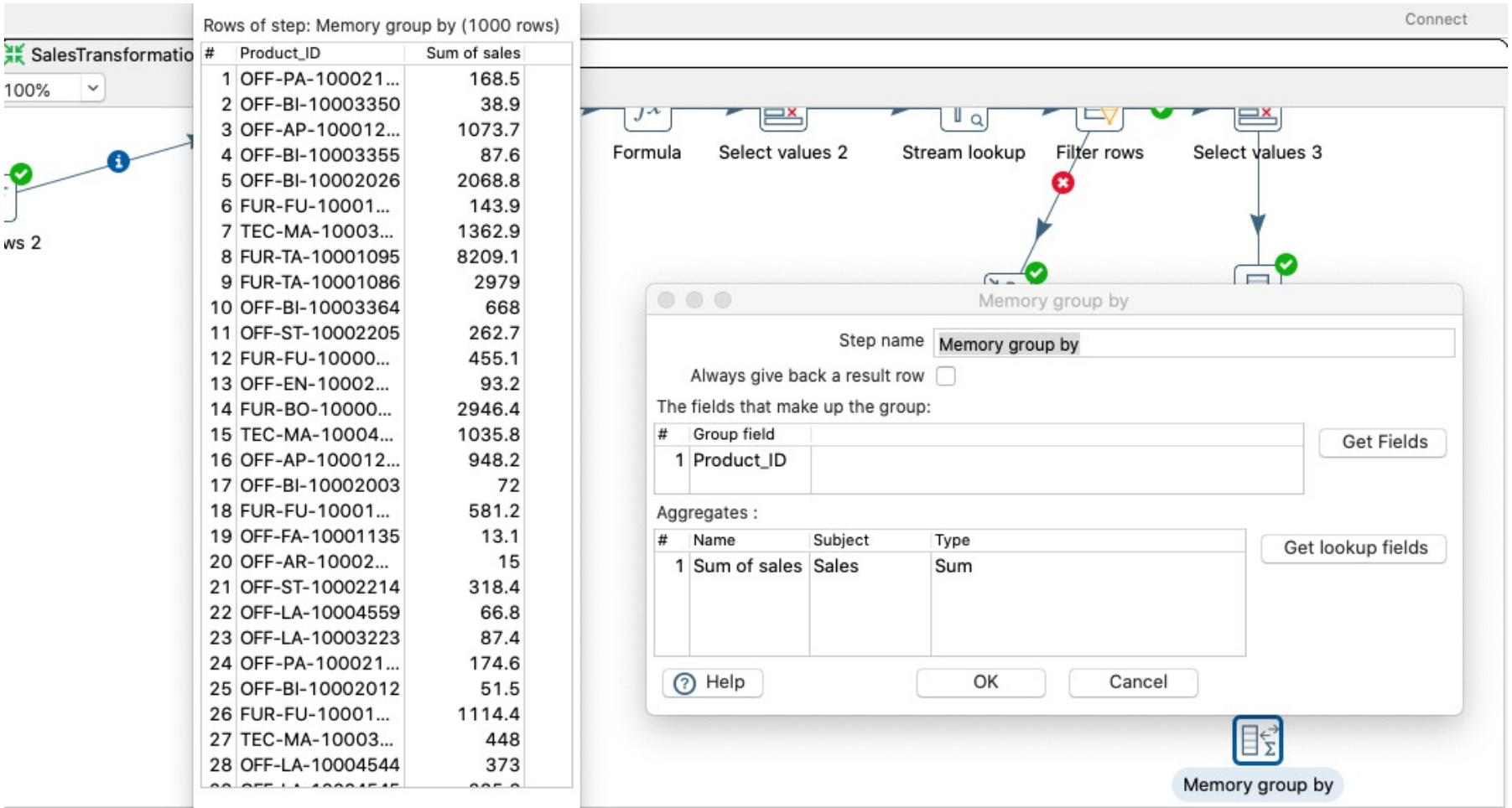
- Total sales for each product
- Total money that each customer has paid
- The number of customers that belong to each State

Data aggregation can be done by using the **group by** step

- Two options:
  - **simple group by**. It need that the field on which we are grouping should be alphabetically sorted
  - **memory group by.** We can use it if the data is not sorted by the grouping field

Transformation: Data aggregation - group by – Sales example

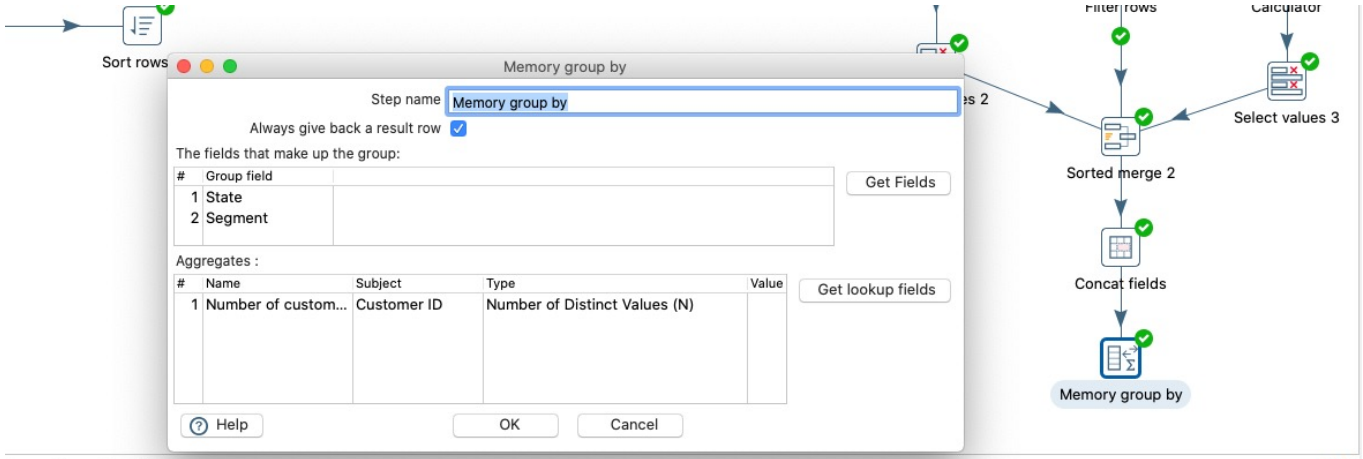
- Total sales for each product (SalesTransformation_clean2_validation2_dates_errorH_Tgroupby.ktr)



Transformation: Data aggregation - group by – Customer example

- How many **customers** belong to each **state** distributed into the three different **segment** (CustomerTransformation_check5_ded clean_fuzzy_Validation_errorh_Tconcatenate_Tgb.ktr)

= For each **state** dividing by **segment**, how many **customers** are there?



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	State	Segment	Number of customers
1	Rhode Island	Corporate	3
2	Colorado	Consumer	11
3	Illinois	Home Office	16
4	Wisconsin	Consumer	4
5	Oklahoma	Consumer	2
6	Iowa	Home Office	1
7	California	Home Office	27
8	Tennessee	Corporate	7
9	Colorado	Home Office	3
10	Louisiana	Home Office	1
11	Pennsylvania	Corporate	18
12	Mississippi	Home Office	1
13	New Mexico	Consumer	2
14	Texas	Home Office	9

Transformation: examples

- Concatenating data
- Data aggregation – group by
- Normalization and denormalization
- Create number ranges

Transformation: Normalizing

Denormalization

Customer segment

States

States	Consumer	Corporate	Home Office
California	86	45	26
New York	44	24	19
Texas	42	25	7
Pennsylvania	23	18	9
Illinois	14	11	16
Ohio	24	9	5
Washington	27	9	2
North Carolina	17	9	4

A pivot table or a cross tabulation of two dimensions

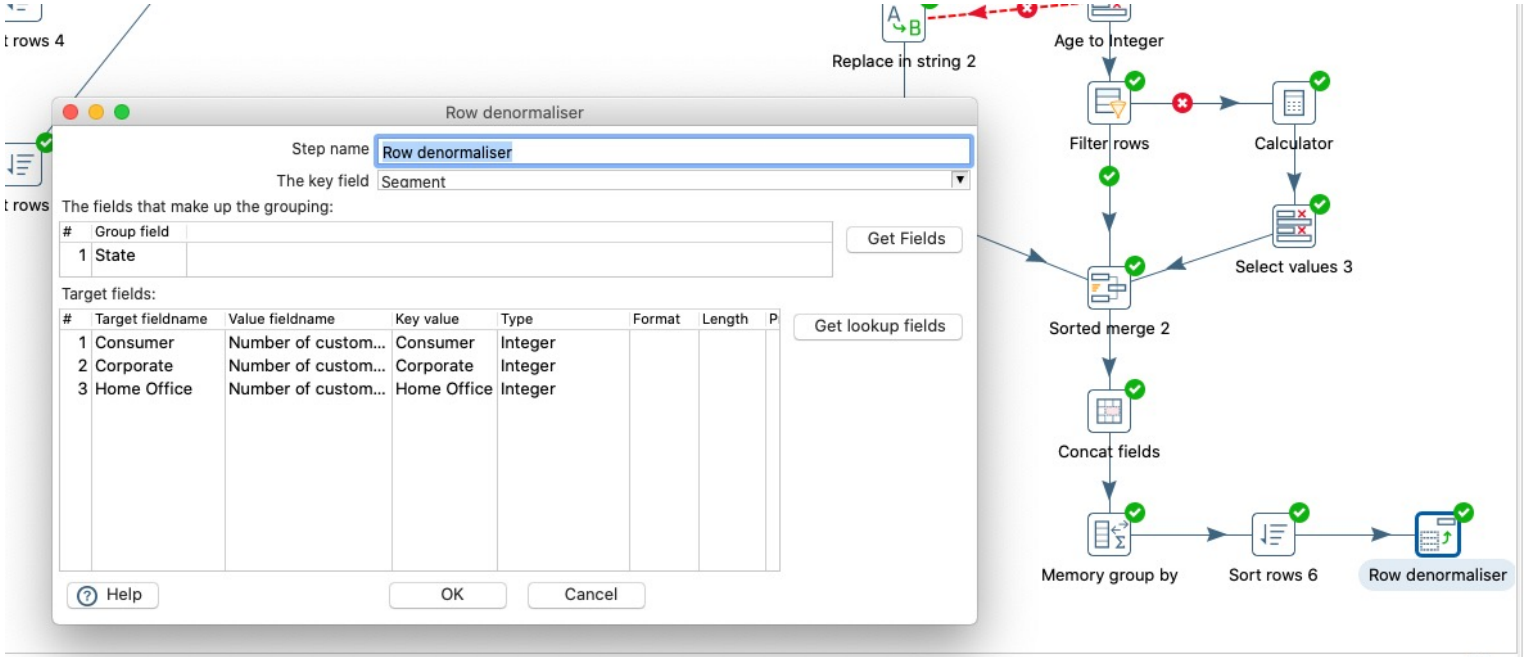
Normalization: Convert columns in rows

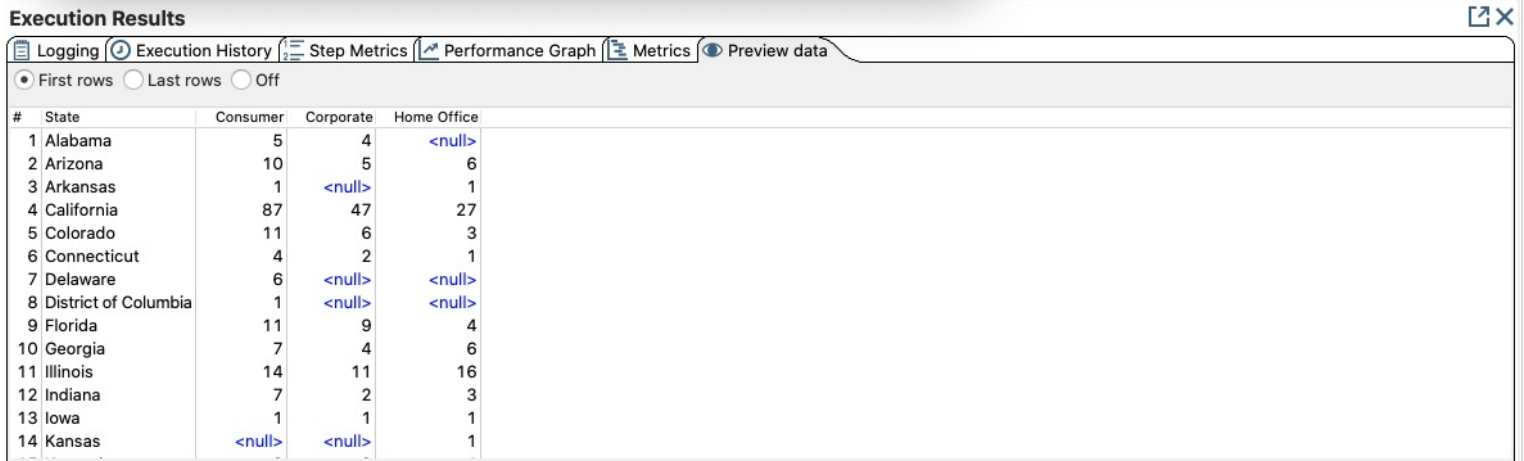
State	Segment	Number of Customers
California	Consumer	86
California	Corporate	45
California	Home Office	26
New York	Consumer	44
New York	Corporate	24
New York	Home Office	19
Texas	Consumer	42
Texas	Corporate	25
Texas	Home Office	7
Pennsylvania	Consumer	23
Pennsylvania	Corporate	18
Pennsylvania	Home Office	9
Illinois	Consumer	14
Illinois	Corporate	11
Illinois	Home Office	16
Ohio	Consumer	24
Ohio	Corporate	9
Ohio	Home Office	5

Transformation: Denormalization – Customer example

- How many customers belong to each state distributed into the three different segment

(CustomerTransformation_check5_dedclean_fuzzy_Validation_errorh_Tconcatenate_Tgb_norm.ktr)





#	State	Consumer	Corporate	Home Office
1	Alabama	5	4	<null>
2	Arizona	10	5	6
3	Arkansas	1	<null>	1
4	California	87	47	27
5	Colorado	11	6	3
6	Connecticut	4	2	1
7	Delaware	6	<null>	<null>
8	District of Columbia	1	<null>	<null>
9	Florida	11	9	4
10	Georgia	7	4	6
11	Illinois	14	11	16
12	Indiana	7	2	3
13	Iowa	1	1	1
14	Kansas	<null>	<null>	1

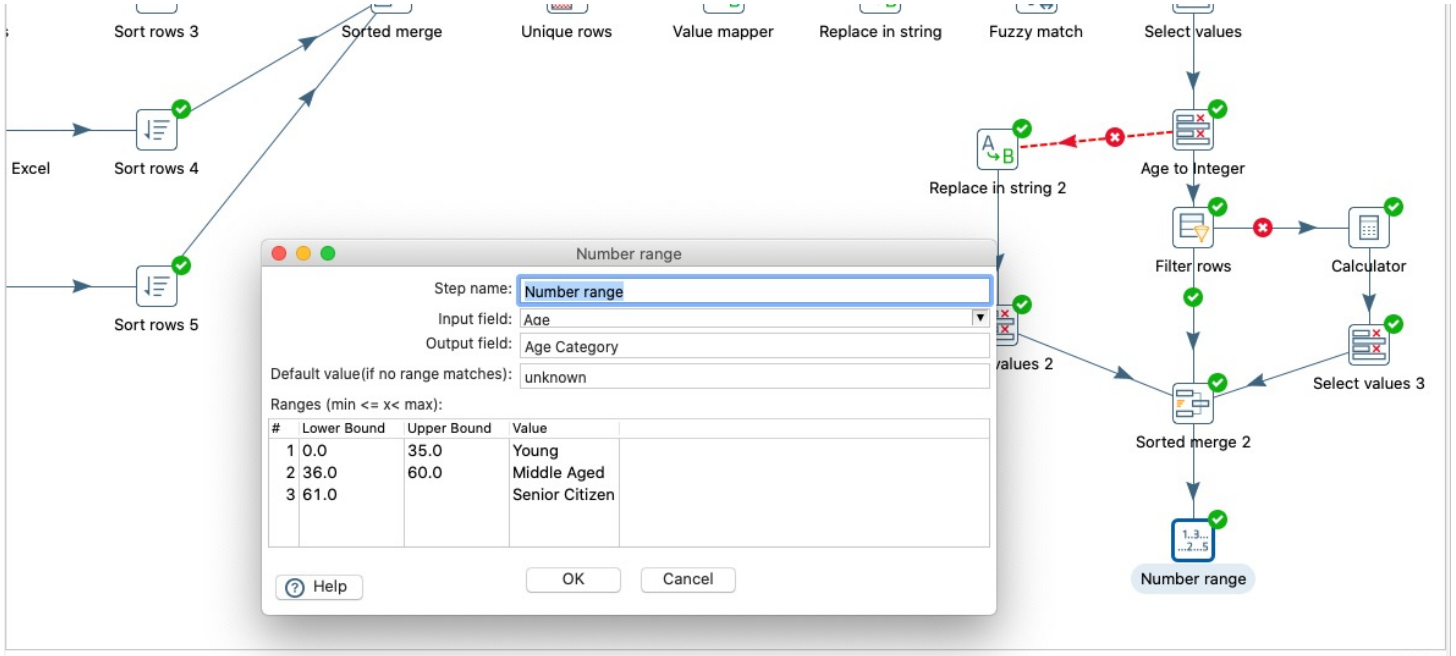
Transformation: examples

- Concatenating data
- Data aggregation – group by
- Normalization and denormalization
- Create number ranges

Transformation: Number Range – Customer data example

- Categorize customers on the basis of their age: young, middle aged, and senior citizen

(CustomerTransformation_check5_ded clean_fuzzy_Validation_errorh_Tnrange.ktr)



Execution Results

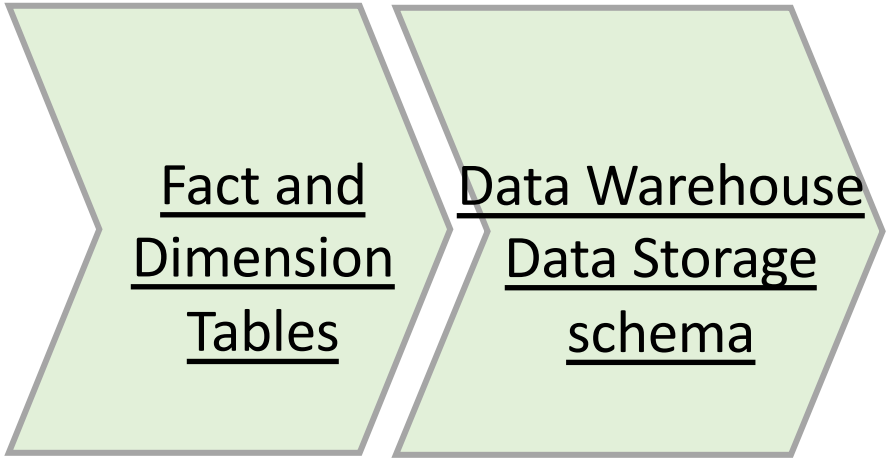
Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows Last rows Off

#	Customer ID	Customer Name	Segment	Age	City	State	Country	Postal Code	Region	Age Category
1	AA-10645	Anna Andreadi	Consumer	32	Chester	Pennsylvania	United States	19013	East	Young
2	AB-10060	Adam Bellavance	Home Office	25	New York City	New York	United States	10009	East	Young
3	AB-10150	Aimee Bixby	Consumer	65	Long Beach	New York	United States	11561	East	Senior Citizen
4	AB-10255	Alejandro Ballentine	Home Office	34	Lorain	Ohio	United States	44052	East	Young
5	AC-10615	Ann Chong	Corporate	61	New York City	New York	United States	10009	East	Senior Citizen
6	AG-10390	Allen Goldenen	Consumer	47	Cincinnati	Ohio	United States	45231	East	Middle Aged
7	AG-10495	Andrew Gjertsen	Corporate	24	Philadelphia	Pennsylvania	United States	19140	East	Young
8	AG-10765	Anthony Garverick	Home Office	40	Philadelphia	Pennsylvania	United States	19120	East	Middle Aged
9	AH-10030	Aaron Hawkins	Corporate	60	Philadelphia	Pennsylvania	United States	19134	East	unknown
10	AH-10075	Adam Hart	Corporate	21	New York City	New York	United States	10011	East	Young
11	AH-10465	Amy Hunt	Consumer	24	New York City	New York	United States	10035	East	Young
12	AH-10690	Anna Haberlin	Corporate	39	New York City	New York	United States	10024	East	Middle Aged
13	AI-10855	Arianne Irving	Consumer	35	Philadelphia	Pennsylvania	United States	19120	East	unknown
14	AJ-10960	Astrea Jones	Consumer	30	Rochester	New York	United States	14609	East	Young

Outline

…


Fact and Dimension Tables
Data Warehouse Data Storage schema

…

Fact and Dimension Tables

Facts and Dimensions

- A **fact table** stores numerical measurements/metrics of the business process as a quantity of anything that can be measured
  - Example: products sold, discounts, taxes, number of invoices
- These measurements are referred to as **facts**.
- Fact table holds the quantitative data

- **Dimension tables** contain the textual descriptors of the business. They helps us describe the attribute of the data which is contained in the fact
  - Typical dimensions are products, time, customers, and regions.
- Dimension table holds the qualitative information

- In our Example:
  - Sales table is a fact table. It contains information of total sales, number of units sold or profit
  - Customer and Product table are Dimension tables

Technical details: Keys in Dimension Table

- Primary key represent the record, this is the business key
  - Example: customer_ID
- A **dimension** table must have a technical key also known as a **surrogate key**.
  - Surrogate keys are always integers
  - The fact table references the surrogate key
- The business key will not available in the fact table but we can find only the surrogate key.
  - Improve storage performance
  - This help also in maintaining privacy for personal information: e.g. a tel num that is used as business key is replaced by a surrogate key
- Dimensions should have a special record for the unavailable data
  - i.e., if you get blank or invalid key, there should be a default value for all attributes
- The **business key/ reference key** is also stored to match the data in the dimension table with the data in the source database.

Technical details: Keys in Dimension Table

In Fact table we will refer to the surrogate key
So, it will be a foreign key

Surrogate Key

Business Key

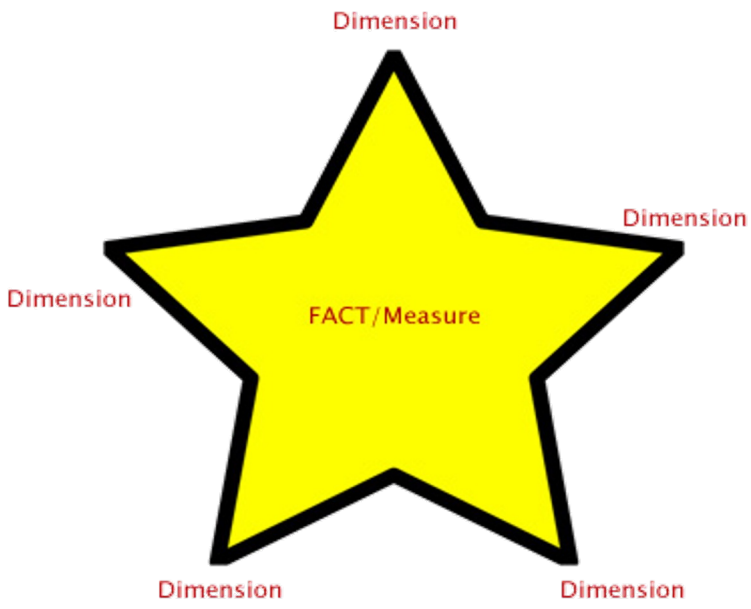
ID	Product ID	Category	Sub-Category	Product Name
1	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase
2	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs Rounded Back
3	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal
4	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table
5	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold N Roll Cart System

Data Warehouse
Data Storage schema

Star Schema vs Snowflake Schema

Star Schema

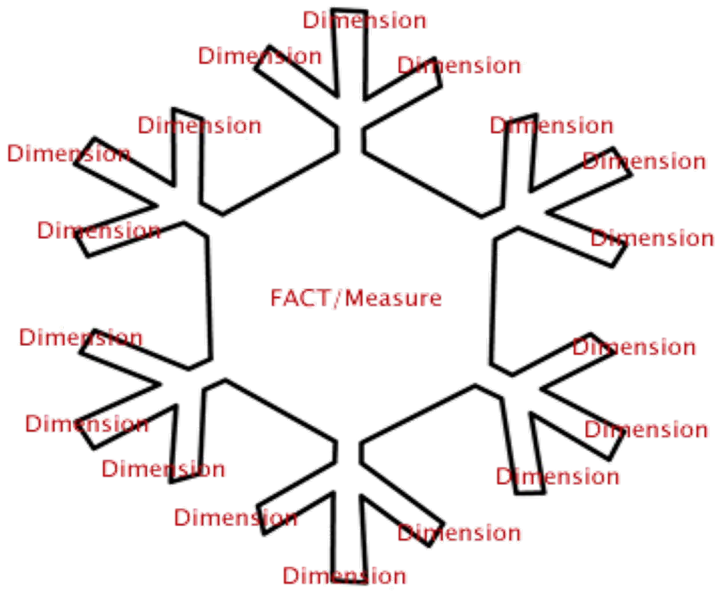
- A **fact table is at the centre** which is connected with several dimension tables



Star Schema

Snowflake Schema

- **Dimension tables are further connected with sub-dimension tables**

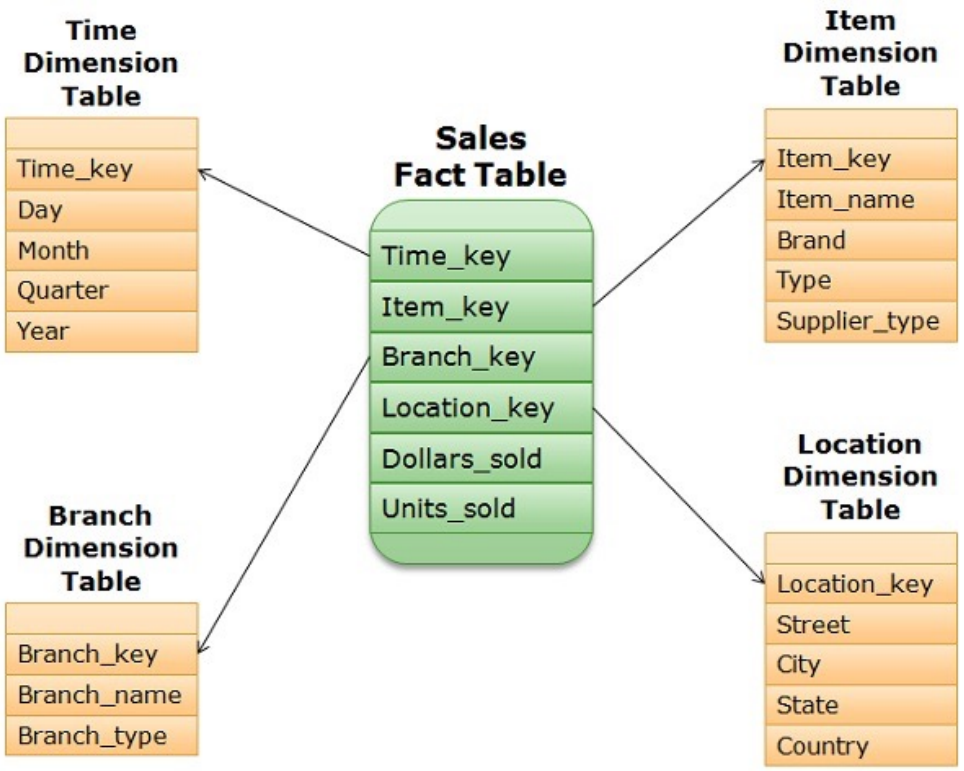


SnowFlake Schema

Star Schema vs Snowflake Schema

Star Schema

- A fact table is at the centre which is connected with several dimension tables



Snowflake Schema

- Dimension tables are further connected with sub-dimension tables