



Splitting a Unified file with multiple years data and loading it in respective tables in Snowflake

This involves below main steps and sub steps:

1. Load unified data from CSV in S3 to Snowflake table
 - Create a table structure in snowflake – using CREATE TABLE

The screenshot shows a Snowflake Data Cloud pipeline with four steps: Start 0, Create Table, S3 Load 0, and a transformation step labeled 'Banking_Txn_filesplit_n_load_Transformation'. Below the pipeline, the 'Create Table' dialog is open, showing the following table properties:

Name	Value	Status
Name	Create Table	OK
Create/Replace	Replace	OK
Database	BANKDATA	OK
Schema	PUBLIC	OK
New Table Name	Banking_data_Unified	OK
Table Type	Permanent	OK
Columns	txn_id, NUMBER, 20, No, No, account_id, NUMBER, 20, No, No, dat...	OK
Default DDL Collation		OK
Primary Keys	txn_id	OK
Clustering Keys		OK
Data Retention Time in D...		OK
Comment		OK

- Use S3 LOAD to load into snowflake

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



Properties | Export | SQL | Help

S3 Load OK

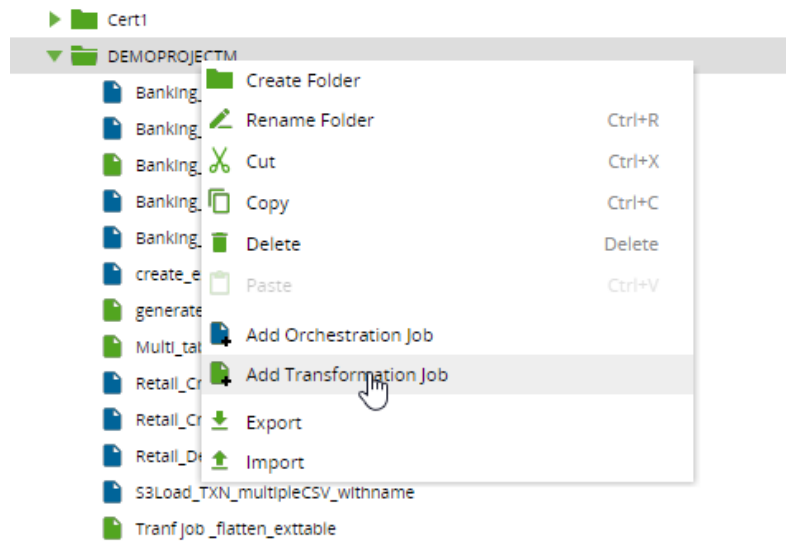
Name	Value	Status
Name	S3 Load 0	OK
Stage	stage_bankingdataren_Txn	OK
Pattern	\${S3_dictionary}/\${Unified_file}.csv	OK
Encryption	None	OK
Warehouse	COMPUTE_WH	OK
Database	BANKDATA	OK
Schema	PUBLIC	OK
Target Table	Banking_data_Unified	OK
Load Columns	txn_id, account_id, date, Type, Opertaion, Amount, Balance, Purp...	OK
Format	[Custom]	OK
File Type	CSV	OK
Compression	AUTO	OK
Record Delimiter		OK
Field Delimiter	,	OK
Skip Header	1	OK
Skip Blank Lines	False	OK
Date Format		OK

- Since we need to loop/iterate YEAR, create a variable for year

Manage Environment Variables						
Name	Type	Behaviour	Matillion_Bankdb	Matillion_dev_env	Matillion_Snowflak...	
example_var	Numeric	Shared		6		
var_env_bank_year	Numeric	Shared	2016	2016		2016

- Creating transformations:
 - Create a transformation Job

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



- Use TABLE INPUT , to bring in data to transform

Name	Value	Status
Name	Banking_data_Unified	OK
Database	BANKDATA	OK
Schema	PUBLIC	OK
Target Table	Banking_data_Unified	OK
Column Names	txn_id, account_id, date, Type, Opertaion, Amount, Balance, Purpo...	OK
Offset		OK

- Use CALCULATOR to add a YEAR column

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



Banking_data_Unified → Calculator 0 → Filter 0 → Rewrite Table 0

Properties | Sample | Metadata | SQL | Plan | Help

Calculator

Name	Value
Name	Calculator 0
Include Input Columns	Yes
Calculations	YEAR("date") , extractedyear

OK

Calculations

Expressions

extractedyear →

1 YEAR("date")

2

Fields Variables

txn_id NUMBER

account_id NUMBER

date DATE

AND OR NOT + - * / || - !-

Functions:

- String Functions
- Math Functions
- Miscellaneous Functions
- Date Functions

- Use FILTER to filter out the year

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



Filter

Name	Value
Name	Filter 0
Filter Conditions	extractedyear, Is, Equal to, \${var_env_bank_year}
Combine Conditions	AND

Filter Conditions

Input Column	Qualifier	Comparator	Value
extractedyear	Is	Equal to	\${var_env_bank_year}

- Write into the respective tables

Rewrite Table

Name	Value
Name	Rewrite Table 0
Warehouse	DEMO_WH_MEDIUM
Database	BANKDATA
Schema	PUBLIC
Target Table	Banking_data_\${var_env_bank_year}
Order By	

- To tie the Orchestration and Iteration Job together we use 'Run transformation' component. But for the variable to loop when transforming we need a LOOP ITERATOR
 - RUN TRANSFORMATION

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



Start 0 → Create Table → S3 Load 0 → Banking_Txn_filesplit_n_load_Transformation

Properties | Export | Help

Run Transformation OK

Name	Value	Status
Name	Banking_Txn_filesplit_n_load_Transformation	OK
Transformation Job	Banking_Txn_filesplit_n_load_Transformation	OK
Set Scalar Variables	Job_var_trans_year, \${var_env_bank_year}	OK
Set Grid Variables		OK

Set Scalar Variables

Variable	Value	Variable Detail
Job_var_trans_year	\${var_env_bank_year}	Defined:
		Type:
		Behaviour:

○ LOOP ITERATOR

Start 0 → Create Table → S3 Load 0 → Banking_Txn_filesplit_n_load_Transformation

Properties | Export | Help

Loop Iterator OK

Name	Value
Name	Loop Iterator 0
Concurrency	Sequential
Variable to Iterate	var_env_bank_year
Starting Value	2016
Increment Value	1
End Value	2021
Break on Failure	No
Record Values In Task His...	Yes
Stop on Condition	No

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



RESULTS:

In mattillion:

<div> <div>Banking_Txn_filesplit_n_load_Transformation</div> <div>Banking_Txn_filesplit_n_load</div> <div>Banking_Txn_filesplit_n_load_SQL</div> <div>Task - Banking_Txn_filesplit_n_load</div> </div>							
✓ Environment: Matillion_Bankdb		Version: default	Queued: 22:55:19		Duration: 31.9s		View Jobs
Job	Component	Duration	Queued	Started	Completed	Row Co...	Message
✓ Banking_Txn_filesplit_n_load		31.9s	22:55:19	22:55:19	22:55:51		
✓ Banking_Txn_filesplit_n_load	Start 0	0.0s	22:55:19	22:55:19	22:55:19		
✓ Banking_Txn_filesplit_n_load	Create Table	1.2s	22:55:19	22:55:19	22:55:20		Created table ["BANKDATA"."PUBLIC"."Banking_data_Un
✓ Banking_Txn_filesplit_n_load	S3 Load 0	9.4s	22:55:20	22:55:20	22:55:29	1048575	
✓ Banking_Txn_filesplit_n_load	Loop Iterator 0	21.2s	22:55:29	22:55:29	22:55:51		6 iterations generated.
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.1s	22:55:29	22:55:29	22:55:32		var_env_bank_year = 2016
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.1s	22:55:29	22:55:29	22:55:32		
✓ Banking_Txn_filesplit_n_load	Banking_data_U...	0.5s	22:55:29	22:55:29	22:55:30	0	
✓ Banking_Txn_filesplit_n_load	Calculator 0	0.1s	22:55:30	22:55:30	22:55:30	0	
✓ Banking_Txn_filesplit_n_load	Filter 0	0.1s	22:55:30	22:55:30	22:55:30	0	
✓ Banking_Txn_filesplit_n_load	Rewrite Table 0	2.4s	22:55:30	22:55:30	22:55:32	28205	
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.4s	22:55:32	22:55:32	22:55:36		var_env_bank_year = 2017
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.4s	22:55:32	22:55:32	22:55:36		
✓ Banking_Txn_filesplit_n_load	Banking_data_U...	1.0s	22:55:32	22:55:32	22:55:33	0	
✓ Banking_Txn_filesplit_n_load	Calculator 0	0.1s	22:55:33	22:55:33	22:55:34	0	
✓ Banking_Txn_filesplit_n_load	Filter 0	0.1s	22:55:34	22:55:34	22:55:34	0	
✓ Banking_Txn_filesplit_n_load	Rewrite Table 0	2.1s	22:55:34	22:55:34	22:55:36	91628	
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.5s	22:55:36	22:55:36	22:55:39		var_env_bank_year = 2018
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.4s	22:55:39	22:55:39	22:55:43		var_env_bank_year = 2019
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.9s	22:55:43	22:55:43	22:55:47		var_env_bank_year = 2020
✓ Banking_Txn_filesplit_n_load	Banking_Txn_fil...	3.9s	22:55:47	22:55:47	22:55:51		var_env_bank_year = 2021

In Snowflake new tables:

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



The screenshot shows the Snowflake web interface. On the left, a list of tables under the 'PUBLIC' schema includes 'Banking_data_2016' through 'Banking_data_2021' and 'Banking_data_Unified'. A pop-up window for 'Banking_data_2016' displays its details: Type is 'Table', Number of rows is '28.2K', Size is '587.0KB', Cluster Key is '—', Owner is 'ACCOUNTADMIN', Created is '49 minutes ago', and Comment is '—'. In the background, a SQL query is visible: `WHERE TABLE_CATALOG= 'BANKDATA' and TABLE_SCHEMA= 'PUBLIC' and table_name LIKE 'Banking_data_%'; ORDER BY 2`. To the right, a table shows row counts for various tables.

TABLE_NAME	ROW_COUNT
Banking_data_2016	28,205
Banking_data_2017	91,628
Banking_data_2018	133,022
Banking_data_2019	196,779
Banking_data_2020	284,409
Banking_data_2021	314,532
Banking_data_Unified	1,048,575

Total rows in each table

The screenshot shows a SQL query in the Snowflake editor and its results. The query is: `SELECT TABLE_CATALOG, TABLE_NAME, ROW_COUNT FROM INFORMATION_SCHEMA.TABLES WHERE TABLE_CATALOG= 'BANKDATA' and TABLE_SCHEMA='PUBLIC' and table_name LIKE 'Banking_data_%'; ORDER BY 2`. The results table below shows the row counts for each table.

TABLE_CATALOG	TABLE_NAME	ROW_COUNT
BANKDATA	Banking_data_2016	28,205
BANKDATA	Banking_data_2017	91,628
BANKDATA	Banking_data_2018	133,022
BANKDATA	Banking_data_2019	196,779
BANKDATA	Banking_data_2020	284,409
BANKDATA	Banking_data_2021	314,532
BANKDATA	Banking_data_Unified	1,048,575

Cross verifying counts in the unified table:

SPLITTING A SINGLE FILE WITH MULTIPLE YEARS DATA INTO MULTIPLE TABLES IN SNOWFLAKE ALONG WITH DATA



BANKDATA.PUBLIC ▾ Settings ▾

```
1 SELECT YEAR("date") as YEAR, COUNT(*) FROM BANKDATA.PUBLIC."Banking_data_Unified"
2 GROUP BY YEAR("date")
3 ORDER BY 1;
```

1
5
5
7

```
3 SELECT TABLE_CATALOG, TABLE_NAME, ROW_COUNT
7 FROM INFORMATION_SCHEMA.TABLES
9 WHERE TABLE_CATALOG= 'BANKDATA' and TABLE_SCHEMA='PUBLIC' and table_name LIKE 'Ba
1 ORDER BY 2
2
```

Results Chart

YEAR	COUNT(*)
2,016	28,205
2,017	91,628
2,018	133,022
2,019	196,779
2,020	284,409
2,021	314,532