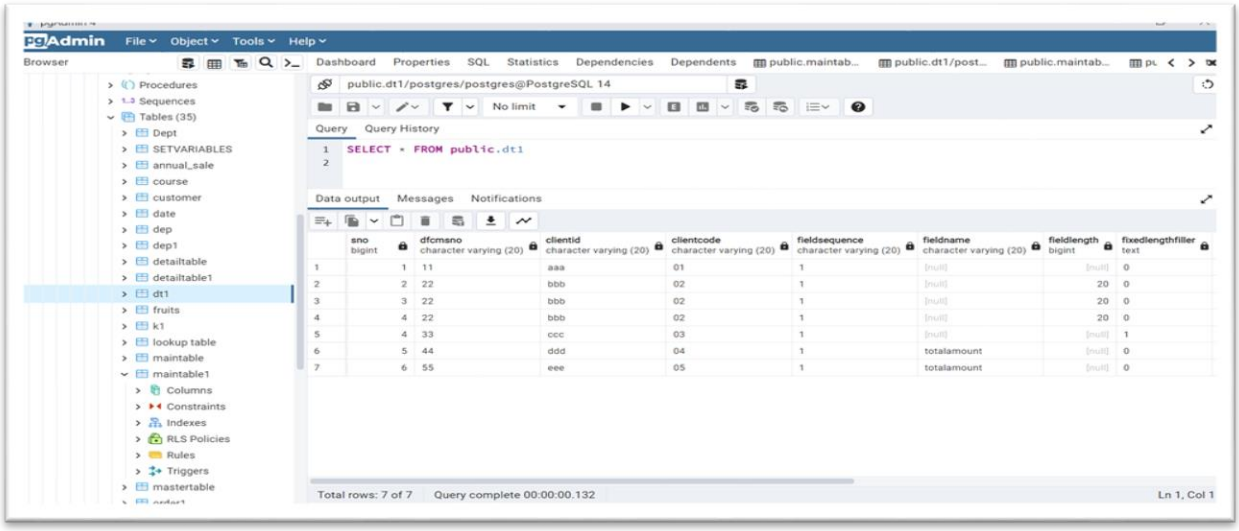
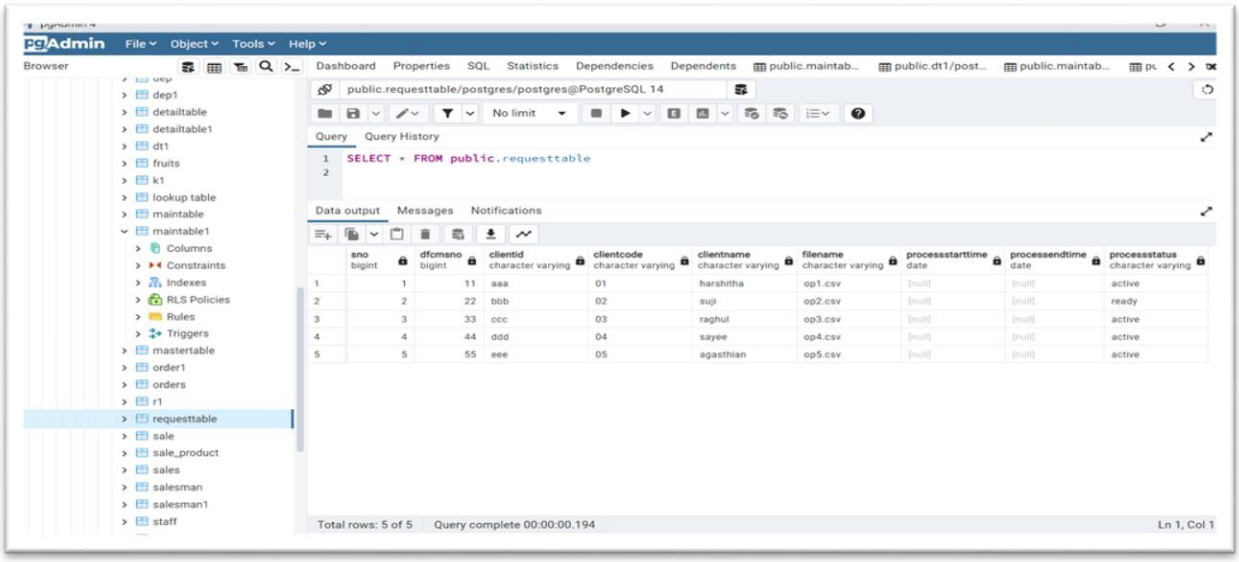
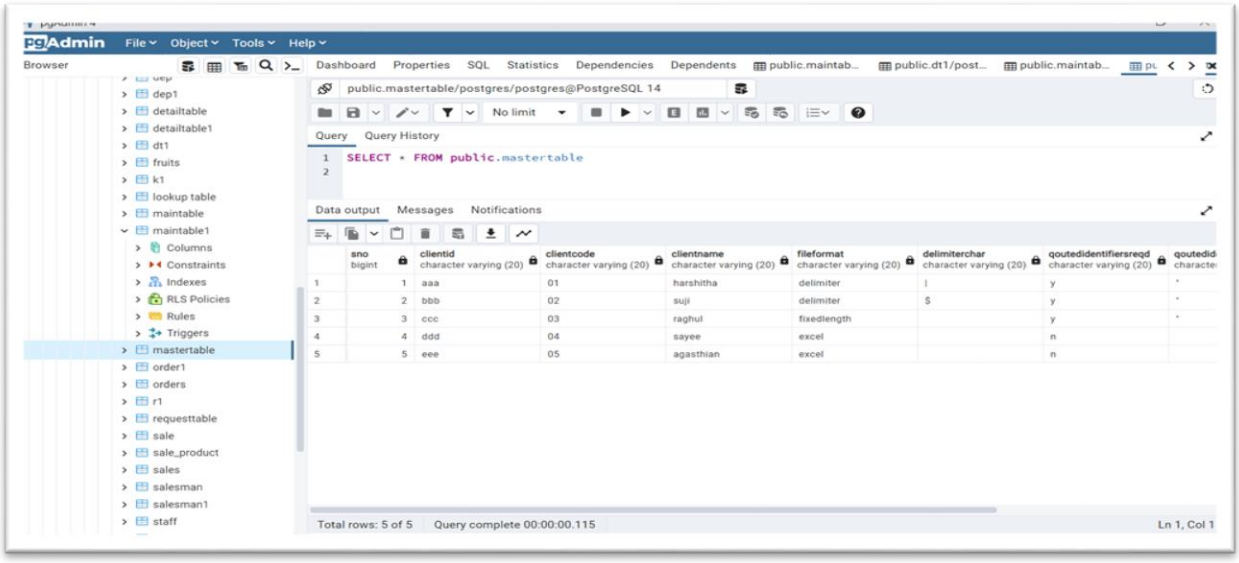
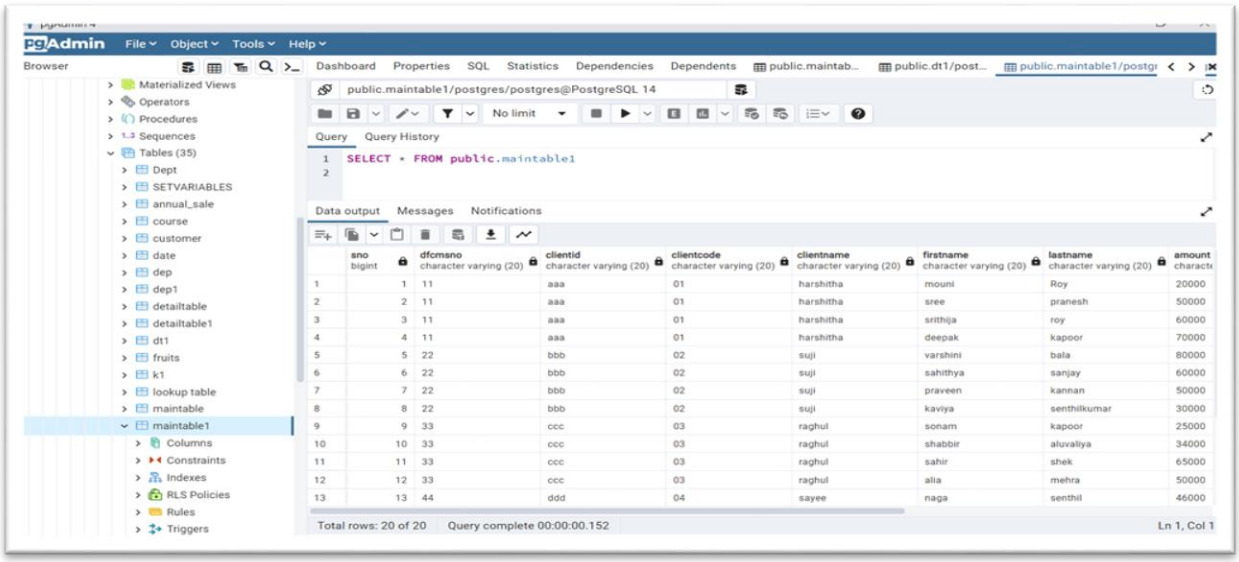
ETL

We have four tables namely Main table, Master table, detail table, request table. Listed below are the tables with dummy data.



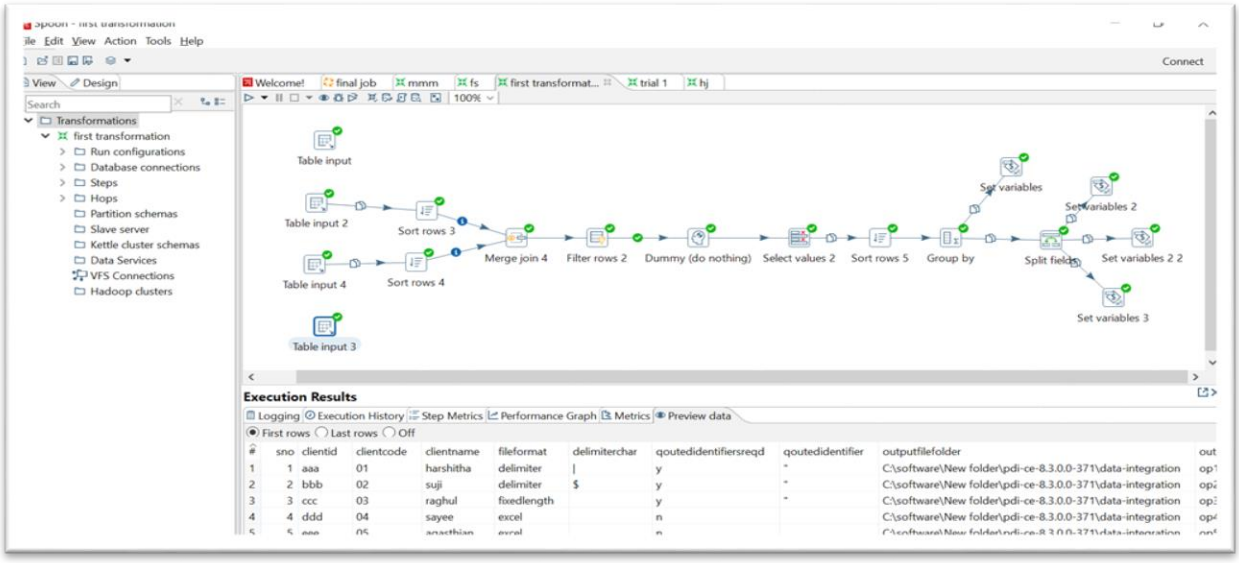


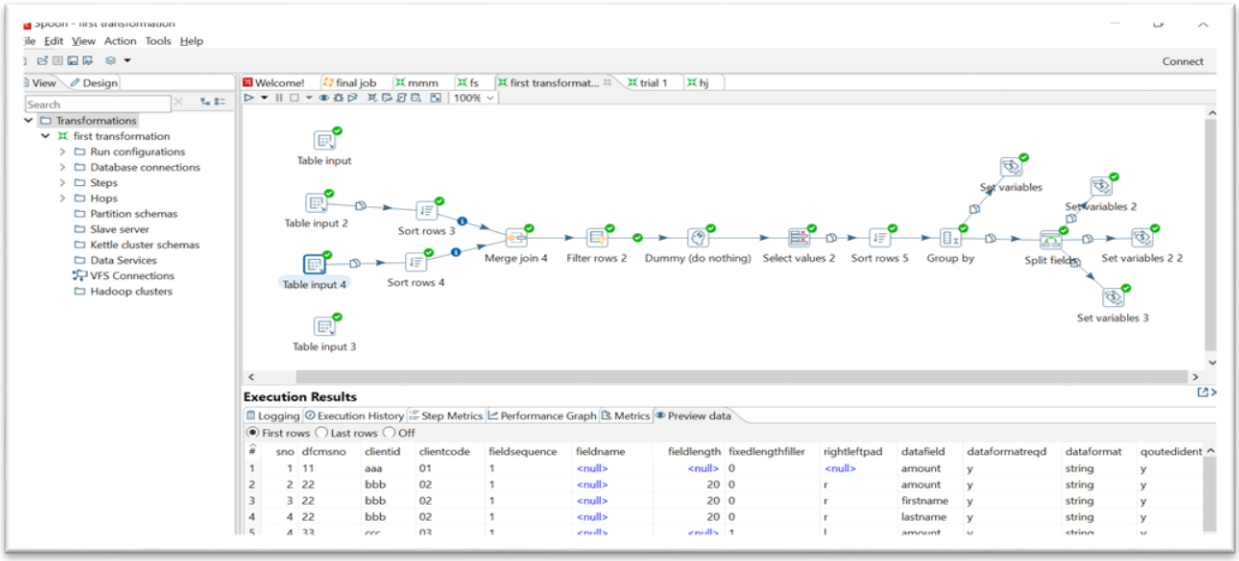


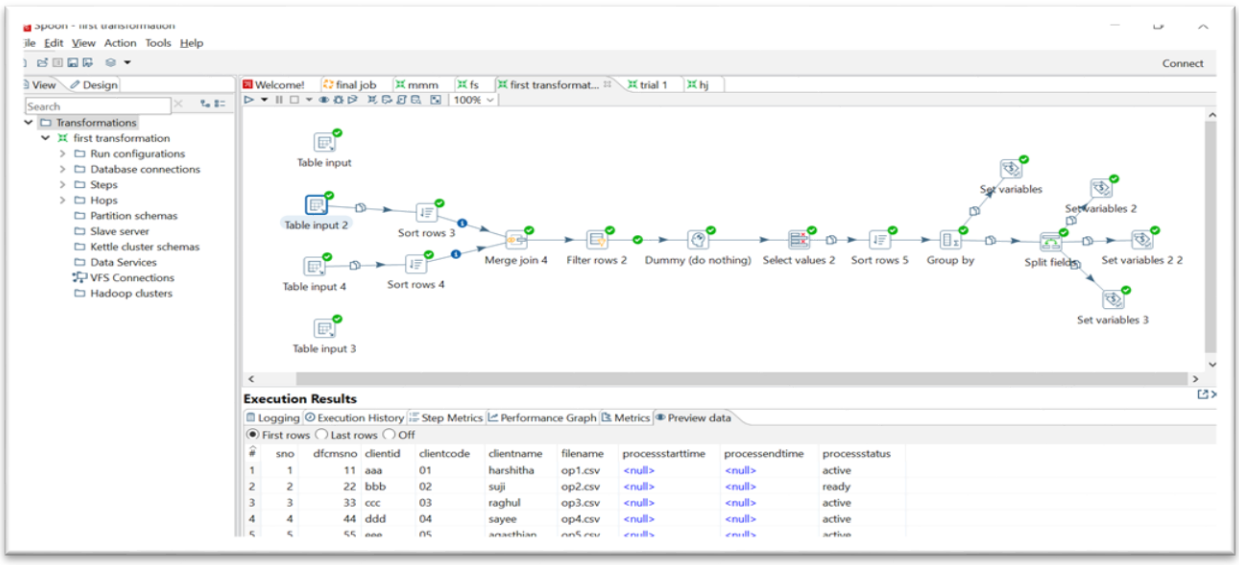


Pentaho Data Integration:

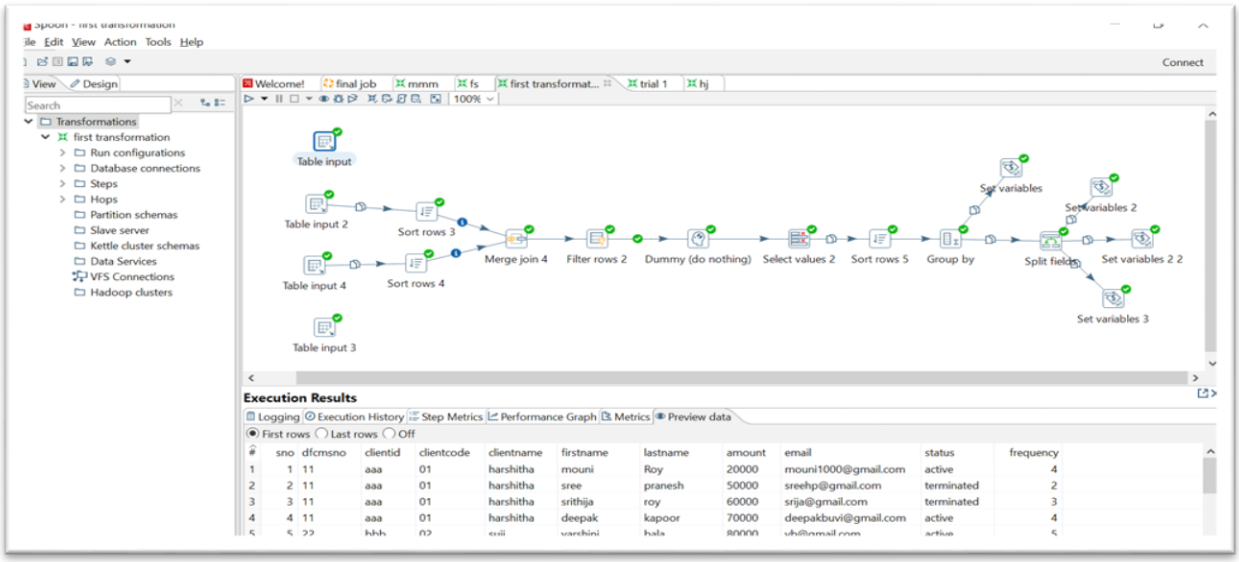
First we fetch those four imported files from PgAdmin to Pentaho tool, Then we Kind of sort and filter it based on our Requirements


File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
first transformation
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Table input 2
Sort rows 3
Table input 4
Sort rows 4
Table input 3
Merge join 4
Filter rows 2
Dummy (do nothing)
Select values 2
Sort rows 5
Group by
Split fields
Set variables
Set variables 2
Set variables 2 2
Set variables 3
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
sno clientid clientcode clientname fileformat delimiterchar qoutedidentifiersreqd qoutedidentifier outputfilefolder
1 1 aaa 01 harshitha delimiter | y " C:\software\New folder\pdi-ce-8.3.0.0-371\data-integration
2 2 bbb 02 suji delimiter $ y " C:\software\New folder\pdi-ce-8.3.0.0-371\data-integration
3 3 ccc 03 raghul fixedlength y " C:\software\New folder\pdi-ce-8.3.0.0-371\data-integration
4 4 ddd 04 sayee excel n C:\software\New folder\pdi-ce-8.3.0.0-371\data-integration

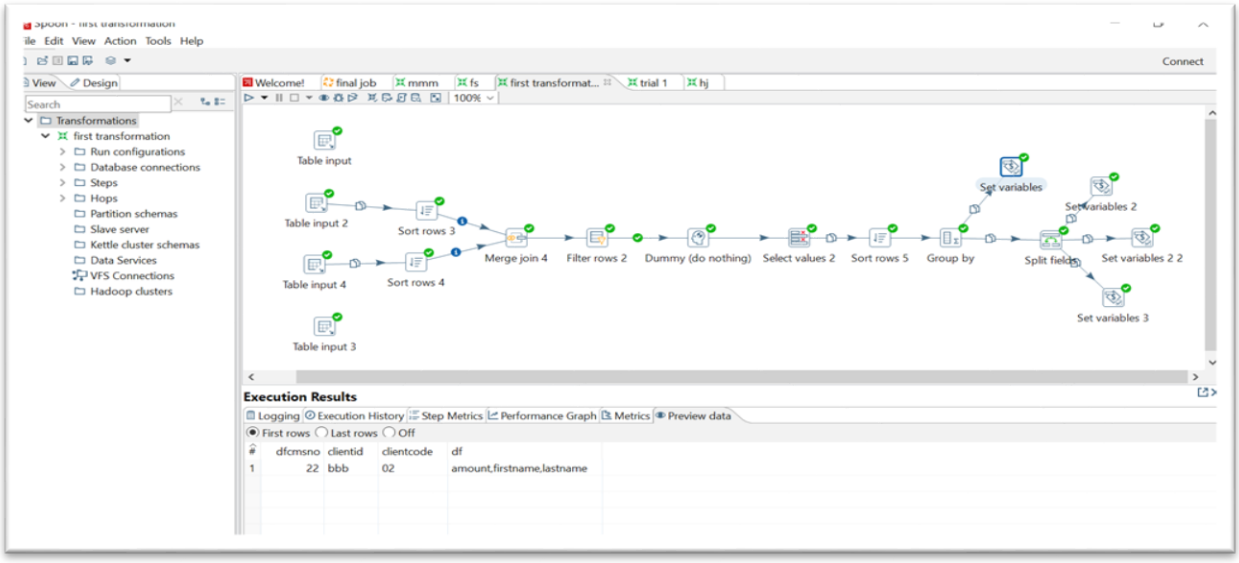

File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
first transformation
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Table input 2
Sort rows 3
Table input 4
Sort rows 4
Table input 3
Merge join 4
Filter rows 2
Dummy (do nothing)
Select values 2
Sort rows 5
Group by
Split fields
Set variables
Set variables 2
Set variables 2 2
Set variables 3
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
sno dfcmsno clientid clientcode fieldsequence fieldname fieldlength fixedlengthfiller rightleftpad datafield dataformatreqd dataformat qoutedident
1 1 11 aaa 01 1 <null> <null> 0 <null> amount y string y
2 2 22 bbb 02 1 <null> 20 0 r amount y string y
3 3 22 bbb 02 1 <null> 20 0 r firstname y string y
4 4 22 bbb 02 1 <null> 20 0 r lastname y string y



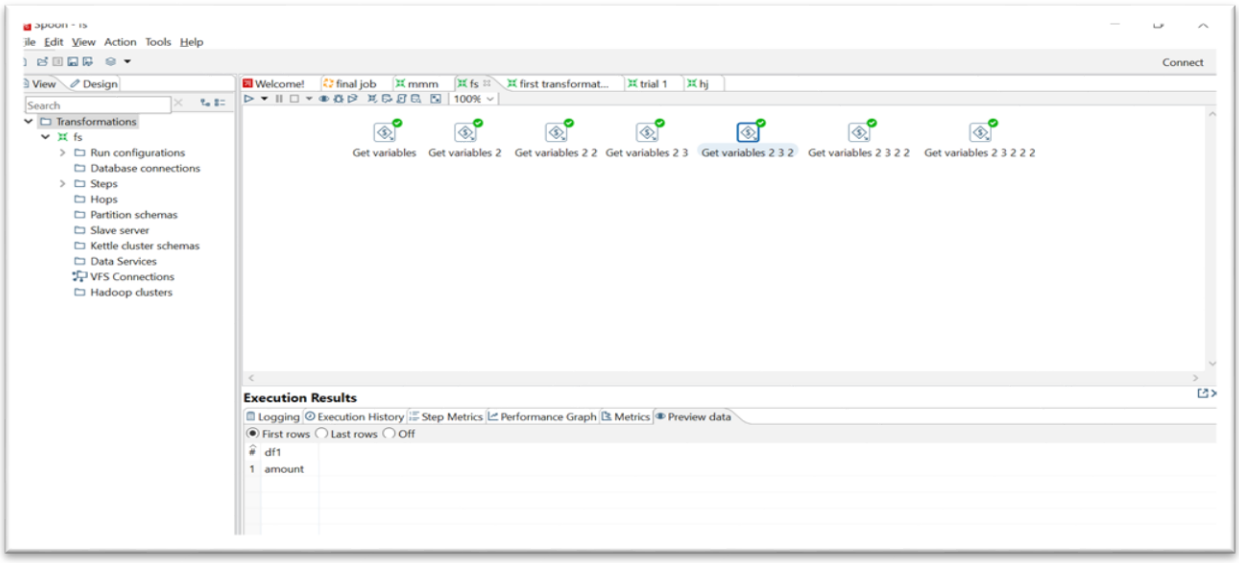
#	sno	dfcmsno	clientid	clientcode	clientname	filename	processstarttime	processendtime	processstatus
1	1	11	aaa	01	harshitha	op1.csv	<null>	<null>	active
2	2	22	bbb	02	suji	op2.csv	<null>	<null>	ready
3	3	33	ccc	03	raghul	op3.csv	<null>	<null>	active
4	4	44	ddd	04	sayee	op4.csv	<null>	<null>	active

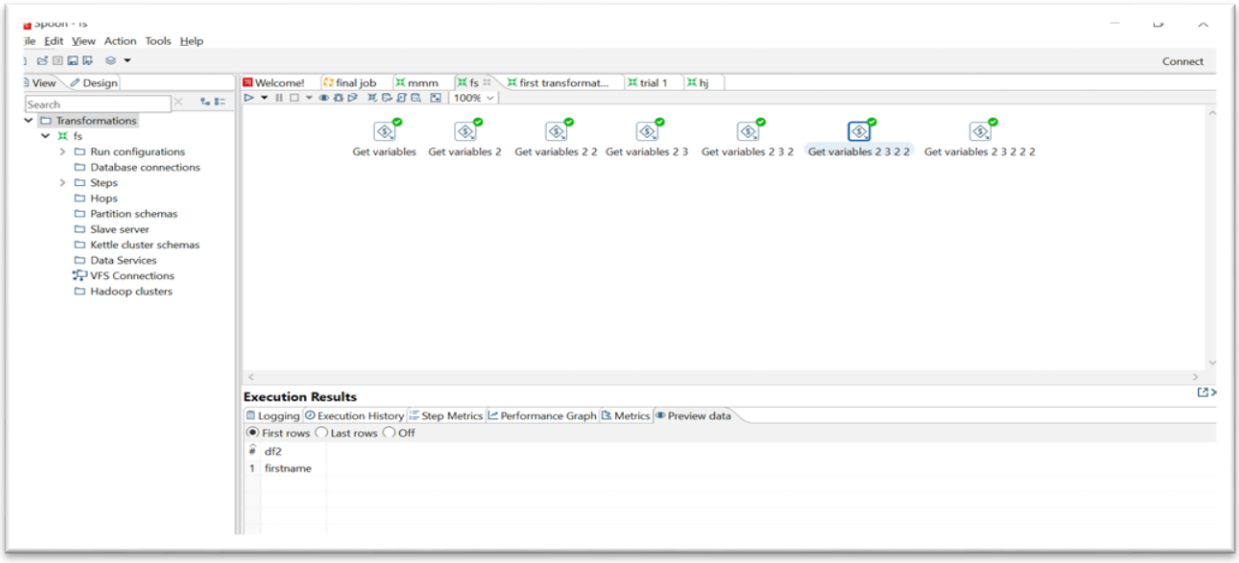


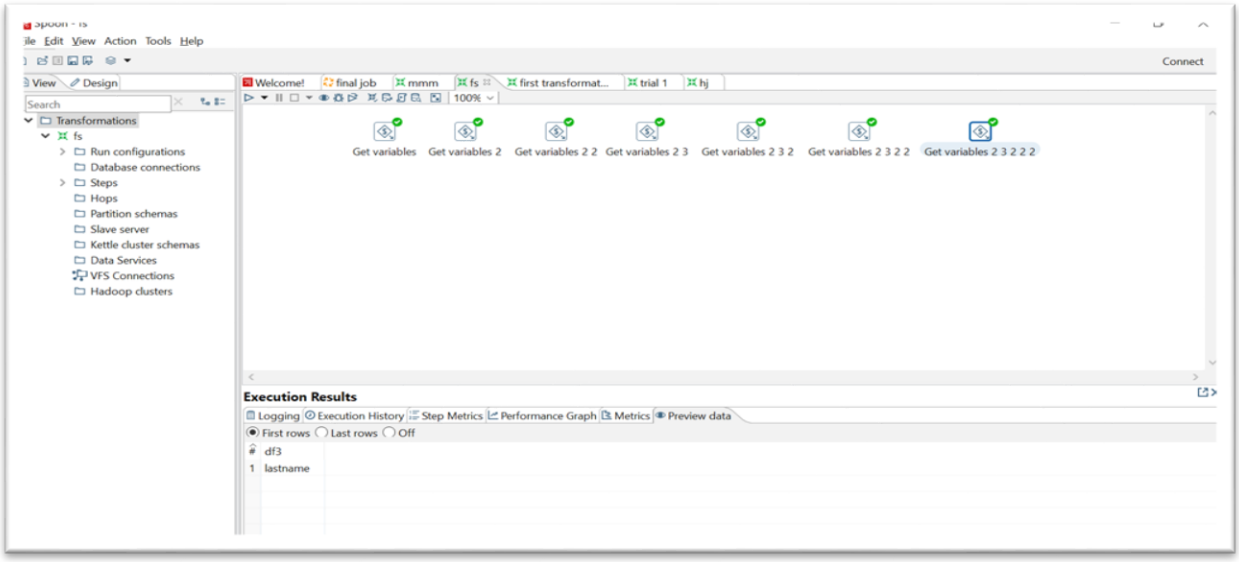
#	sno	dfcmsno	clientid	clientcode	clientname	firstname	lastname	amount	email	status	frequency
1	1	11	aaa	01	harshitha	mouni	Roy	20000	mouni1000@gmail.com	active	4
2	2	11	aaa	01	harshitha	sree	pranesh	50000	sreehp@gmail.com	terminated	2
3	3	11	aaa	01	harshitha	srithija	roy	60000	srija@gmail.com	terminated	3
4	4	11	aaa	01	harshitha	deepak	kapoor	70000	deepakbuvi@gmail.com	active	4



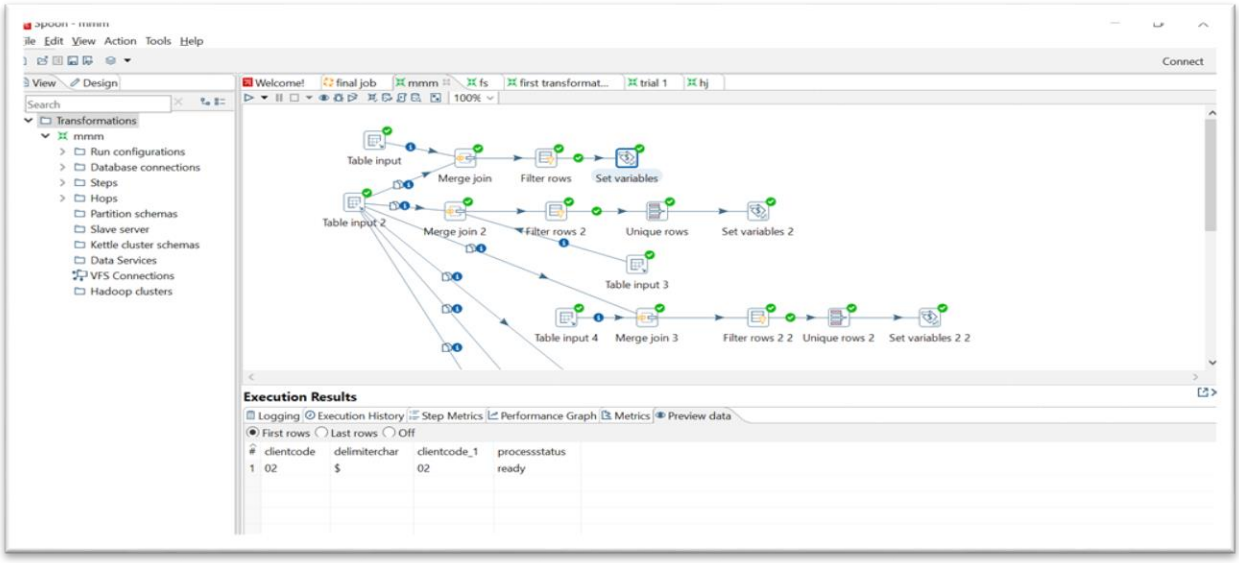
Now we are getting separate variables to pull those data field column in every other part of the transformation

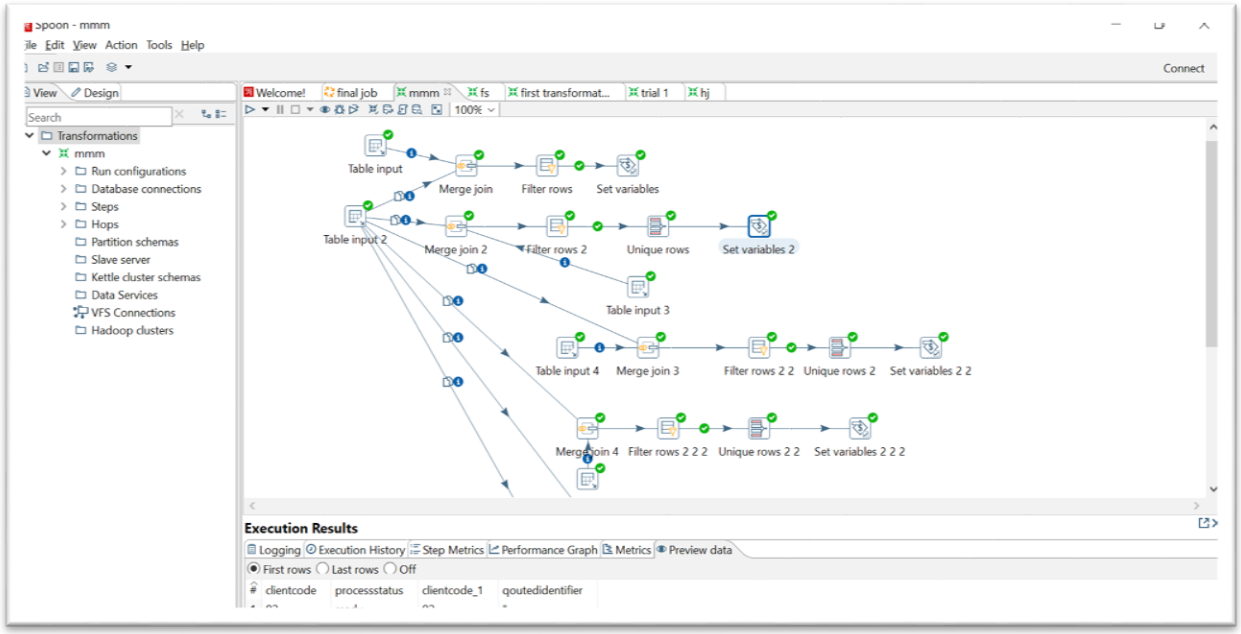


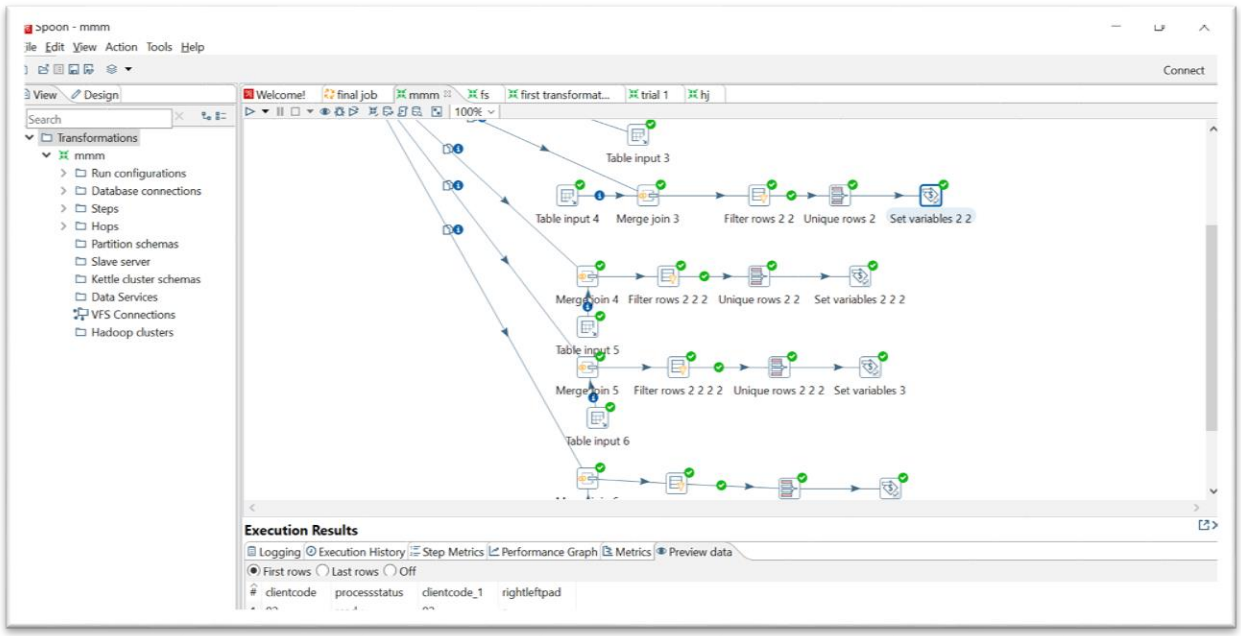


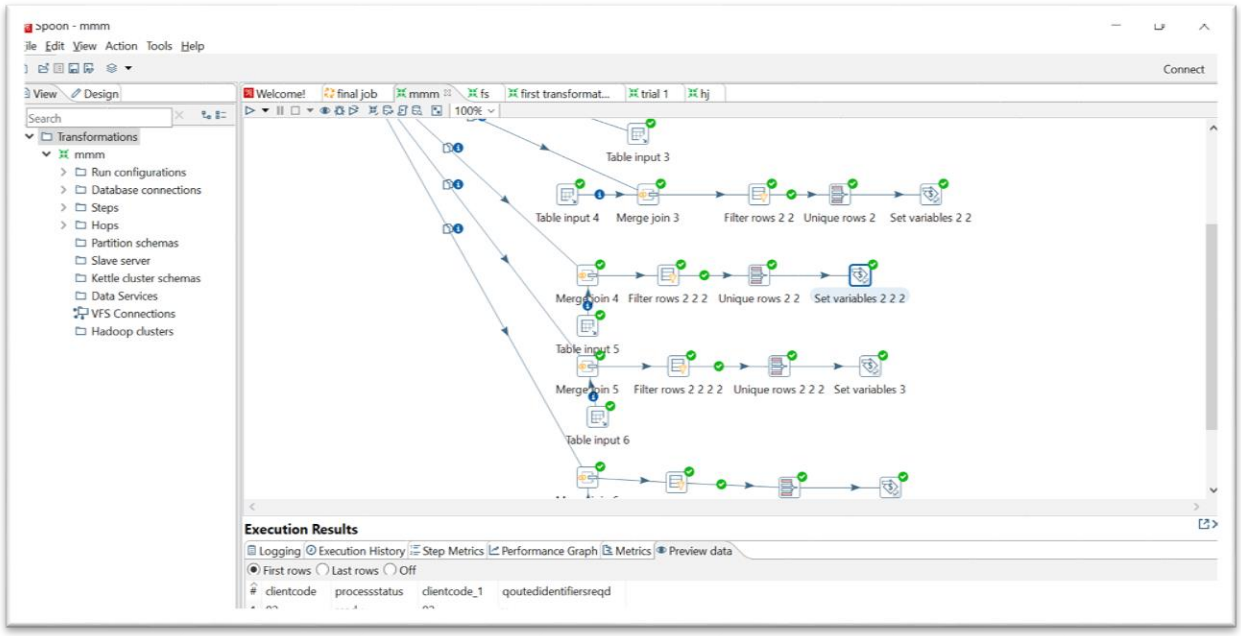


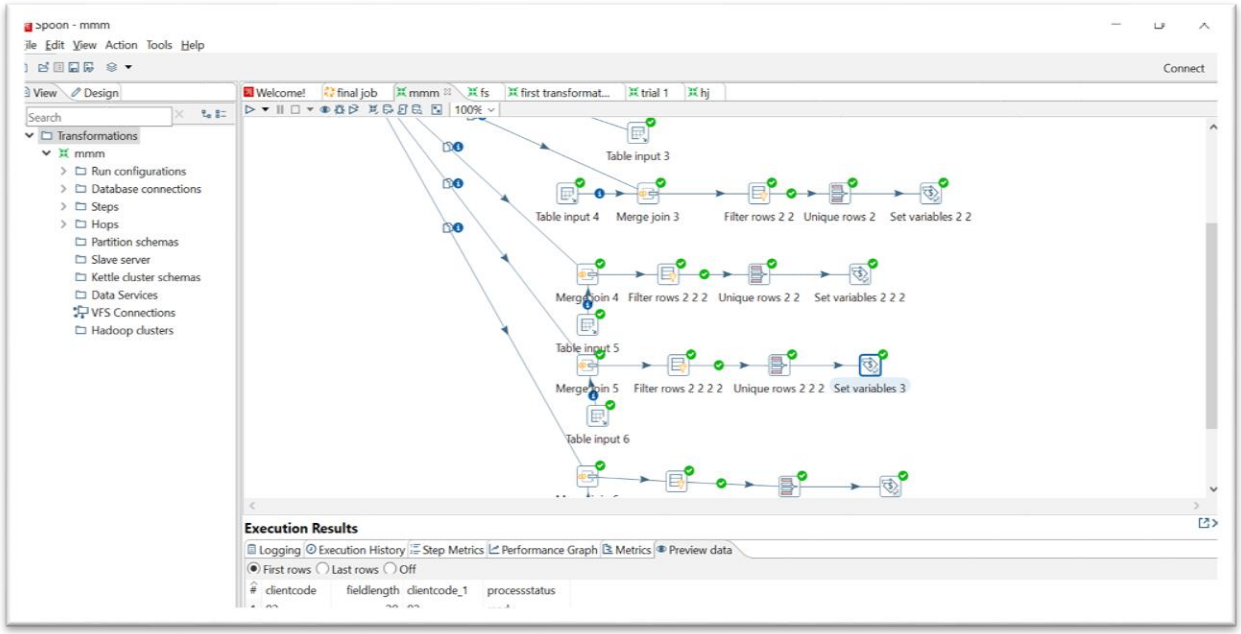
Then we are setting variables to pull certain requirement related data like padding, Fixed length fillers, qouted identifiers required or not etc

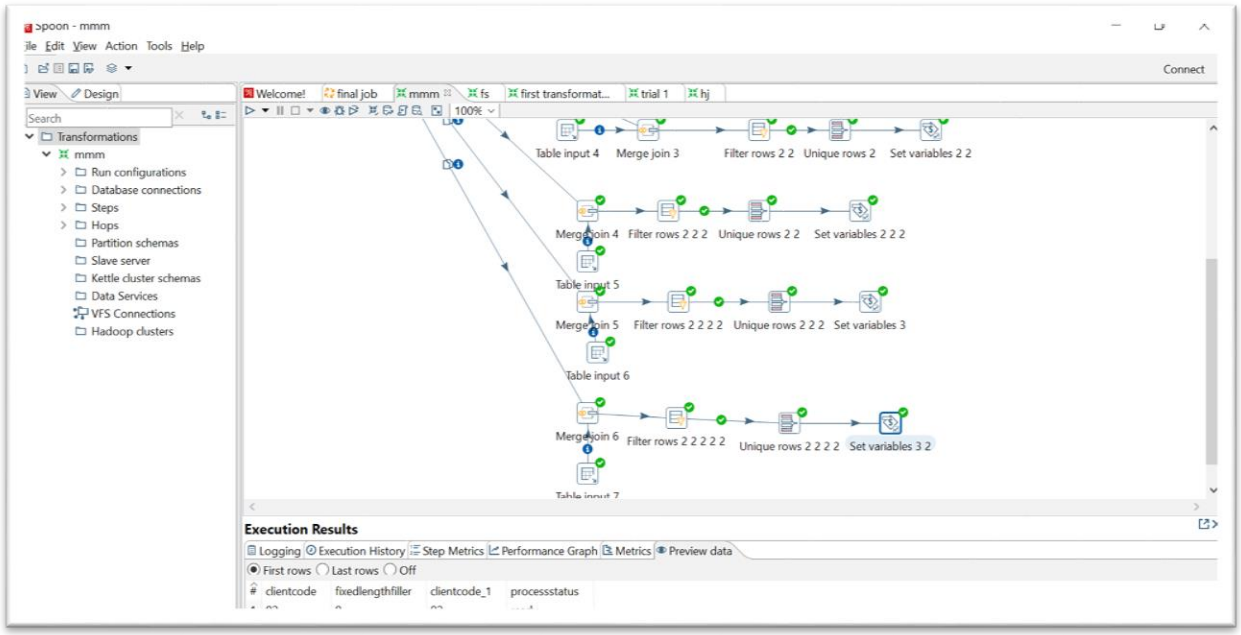

File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
mmm
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Merge join
Filter rows
Set variables
Table input 2
Merge join 2
Filter rows 2
Unique rows
Set variables 2
Table input 3
Table input 4
Merge join 3
Filter rows 2 2
Unique rows 2
Set variables 2 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
clientcode delimiterchar clientcode_1 processstatus
1 02 $ 02 ready


Spoon - mmm
File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
mmm
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Merge join
Filter rows
Set variables
Table input 2
Merge join 2
Filter rows 2
Unique rows
Set variables 2
Table input 3
Table input 4
Merge join 3
Filter rows 2 2
Unique rows 2
Set variables 2 2
Merge join 4
Filter rows 2 2 2
Unique rows 2 2
Set variables 2 2 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
clientcode processstatus clientcode_1 qoutedidentifier

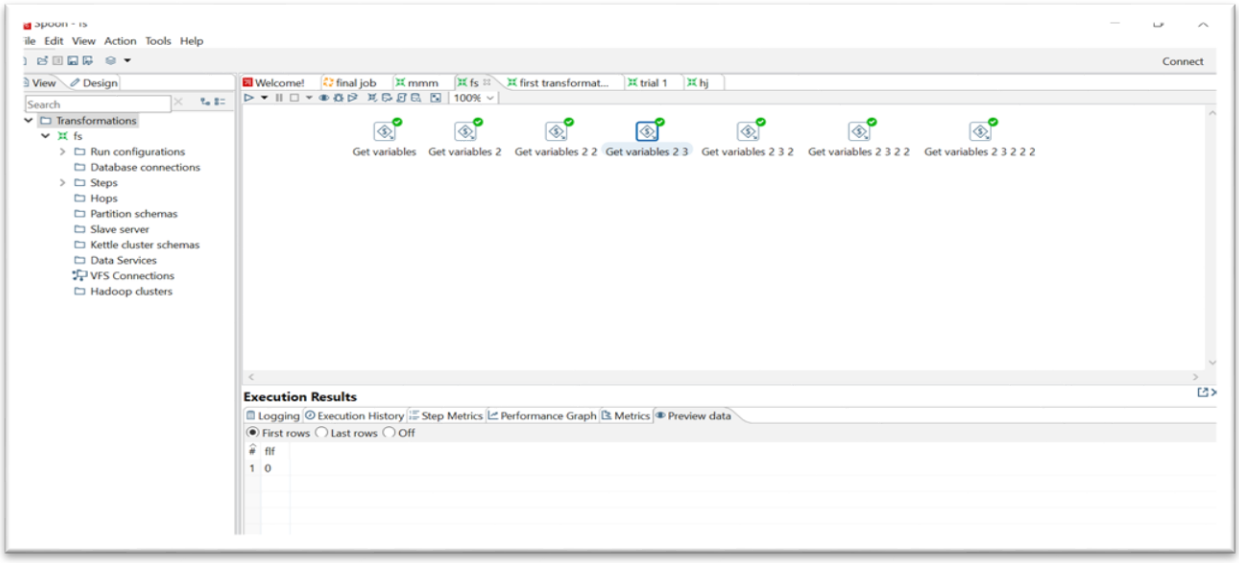

spoon - mmm
File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
mmm
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input 3
Table input 4
Merge join 3
Filter rows 2 2
Unique rows 2
Set variables 2 2
Merge join 4
Filter rows 2 2 2
Unique rows 2 2
Set variables 2 2 2
Table input 5
Merge join 5
Filter rows 2 2 2 2
Unique rows 2 2 2
Set variables 3
Table input 6
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
clientcode
processstatus
clientcode_1
rightleftpad

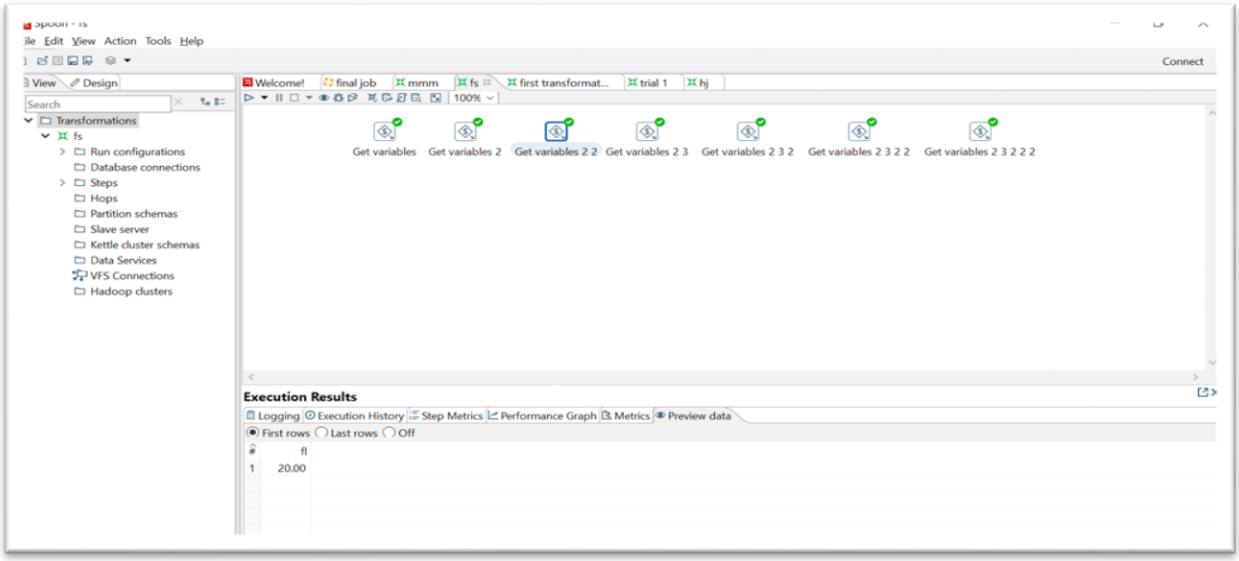

spoon - mmm
File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
mmm
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input 3
Table input 4
Merge join 3
Filter rows 2 2
Unique rows 2
Set variables 2 2
Merge join 4
Filter rows 2 2 2
Unique rows 2 2
Set variables 2 2 2
Table input 5
Merge join 5
Filter rows 2 2 2 2
Unique rows 2 2 2
Set variables 3
Table input 6
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
clientcode
processstatus
clientcode_1
qoutedidentifiersreqd

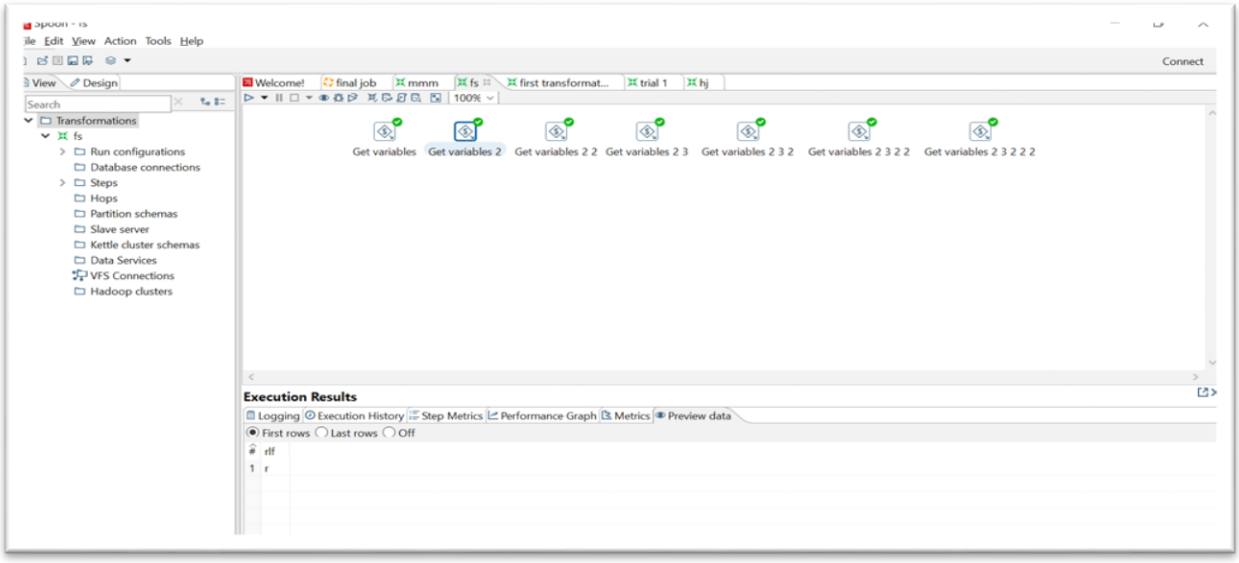


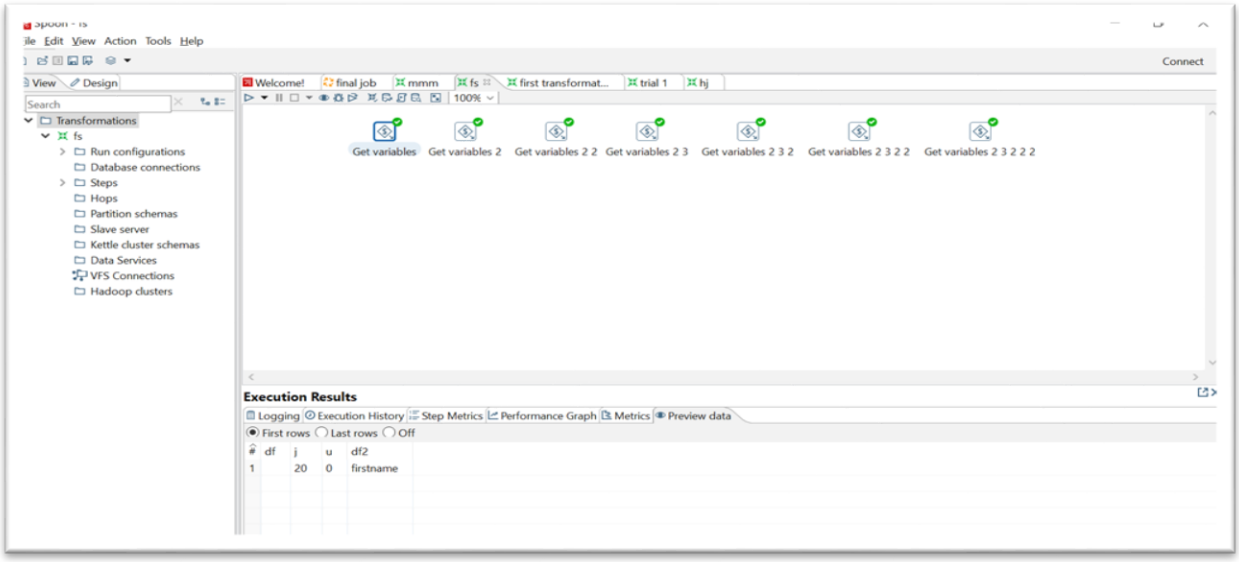


Now we have to get those variables to fetch those data in columns automatically

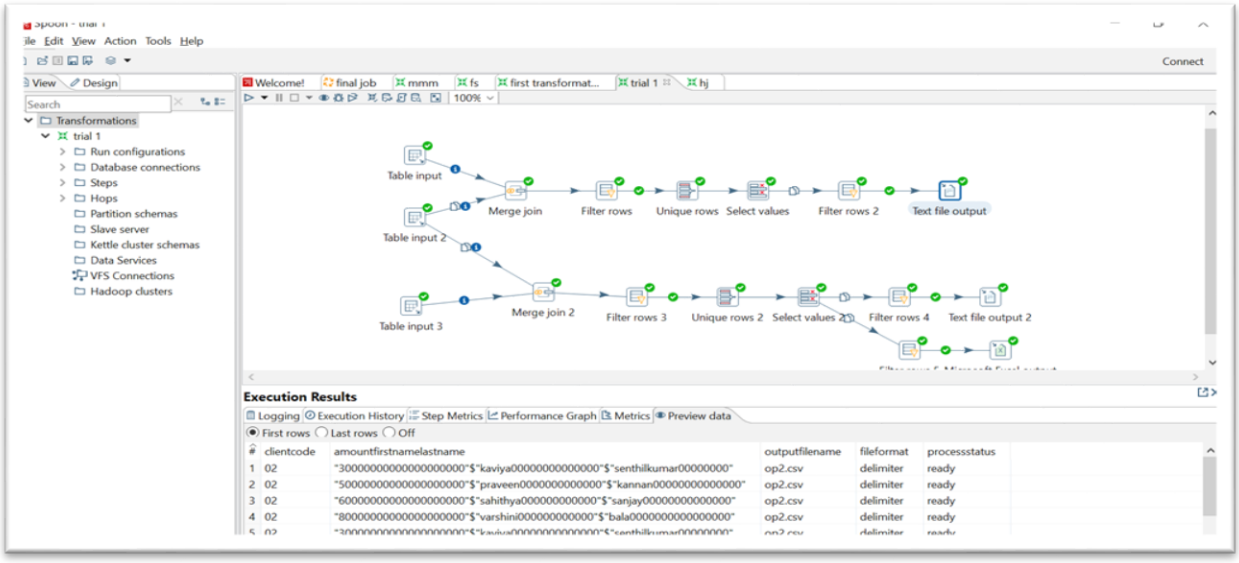

File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
fs
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Get variables
Get variables 2
Get variables 2 2
Get variables 2 3
Get variables 2 3 2
Get variables 2 3 2 2
Get variables 2 3 2 2 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
flf
1 0

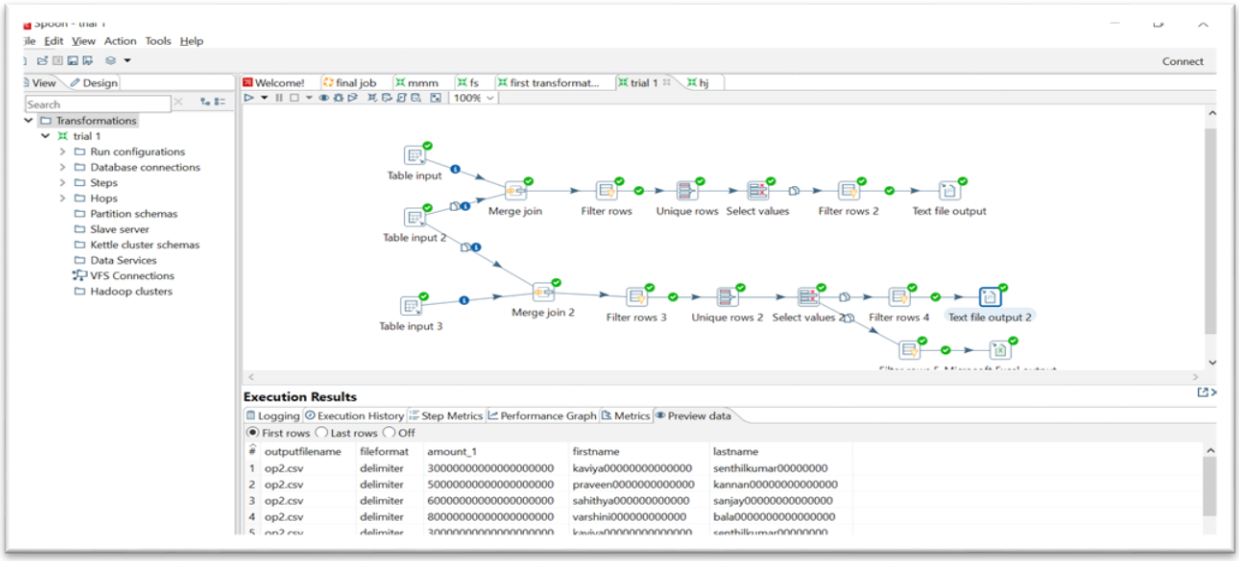

File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
fs
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Get variables
Get variables 2
Get variables 2 2
Get variables 2 3
Get variables 2 3 2
Get variables 2 3 2 2
Get variables 2 3 2 2 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
fl
1 20.00

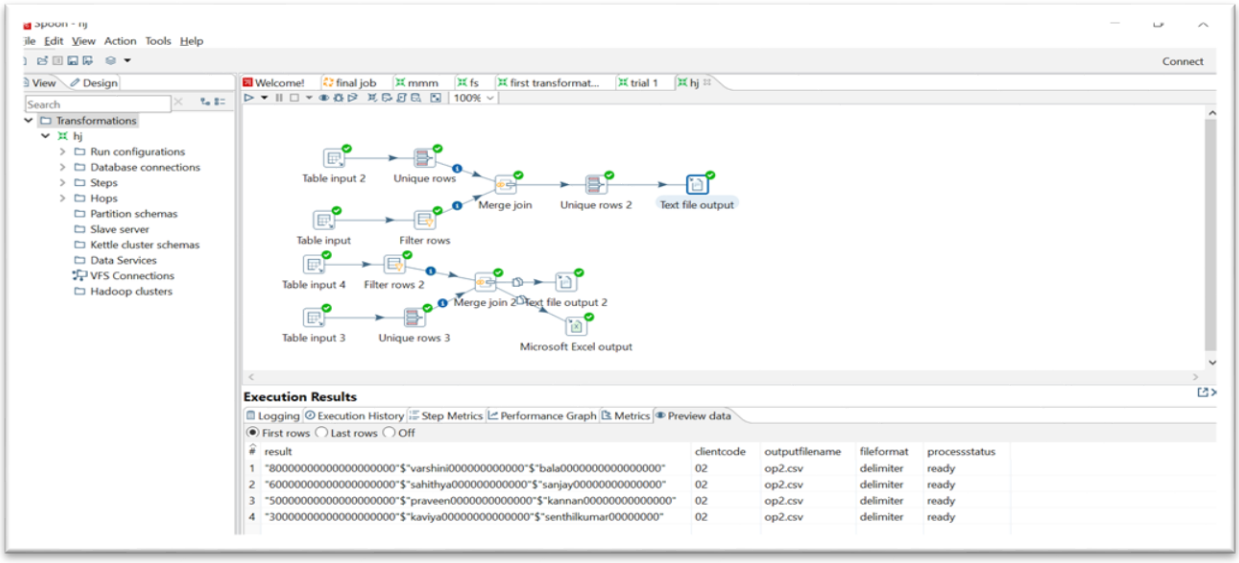




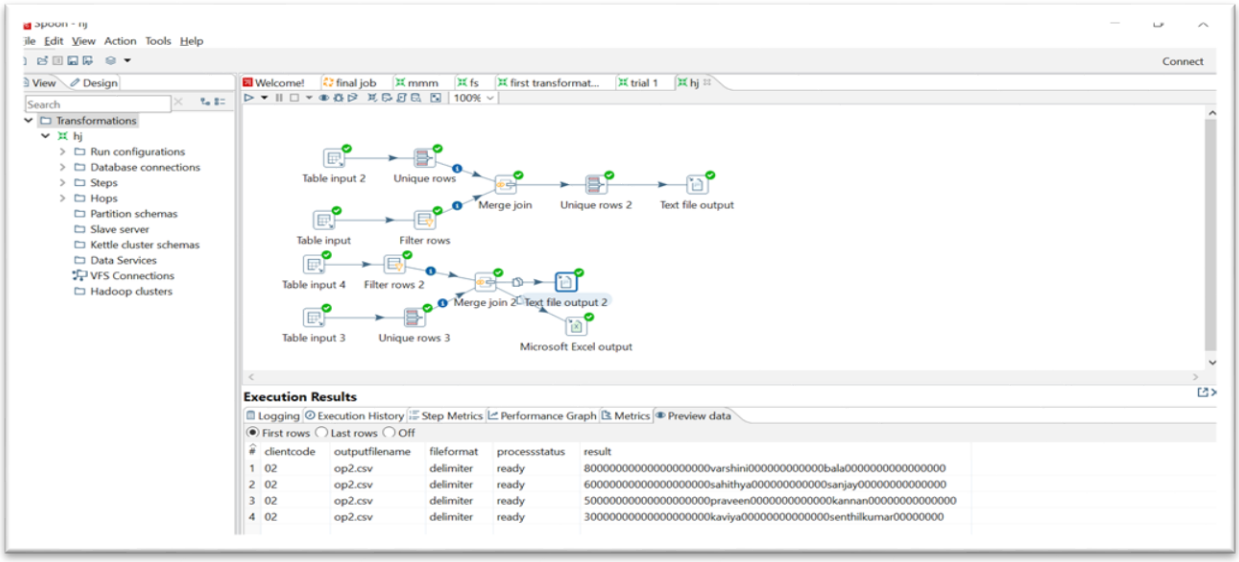
Now we do further steps to bring those data field in the required format and file types


File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
trial 1
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Table input 2
Table input 3
Merge join
Merge join 2
Filter rows
Filter rows 3
Unique rows
Unique rows 2
Select values
Select values 2
Filter rows 2
Filter rows 4
Text file output
Text file output 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
clientcode amountfirstnamelastname outputfilename fileformat processstatus
1 02 "300000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000" op2.csv delimiter ready
2 02 "500000000000000000000"$"praveen00000000000000"$"kannan00000000000000" op2.csv delimiter ready
3 02 "600000000000000000000"$"sahithya000000000000"$"sanjay00000000000000" op2.csv delimiter ready
4 02 "800000000000000000000"$"varshini000000000000"$"bala00000000000000000" op2.csv delimiter ready

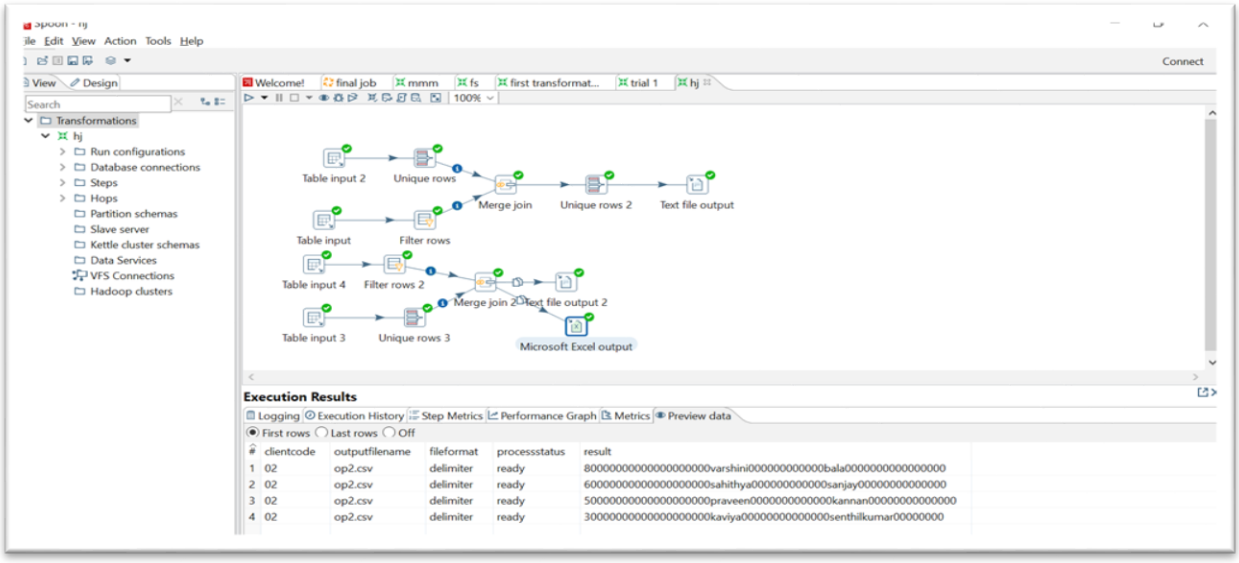

File Edit View Action Tools Help
Connect
View
Design
Search
Transformations
trial 1
Run configurations
Database connections
Steps
Hops
Partition schemas
Slave server
Kettle cluster schemas
Data Services
VFS Connections
Hadoop clusters
Welcome!
final job
mmm
fs
first transformat...
trial 1
hj
100%
Table input
Table input 2
Table input 3
Merge join
Merge join 2
Filter rows
Filter rows 3
Unique rows
Unique rows 2
Select values
Select values 2
Filter rows 2
Filter rows 4
Text file output
Text file output 2
Execution Results
Logging
Execution History
Step Metrics
Performance Graph
Metrics
Preview data
First rows
Last rows
Off
outputfilename fileformat amount_1 firstname lastname
1 op2.csv delimiter 300000000000000000000 kaviya00000000000000 senthilkumar00000000
2 op2.csv delimiter 500000000000000000000 praveen00000000000000 kannan00000000000000
3 op2.csv delimiter 600000000000000000000 sahithya000000000000 sanjay00000000000000
4 op2.csv delimiter 800000000000000000000 varshini000000000000 bala00000000000000000



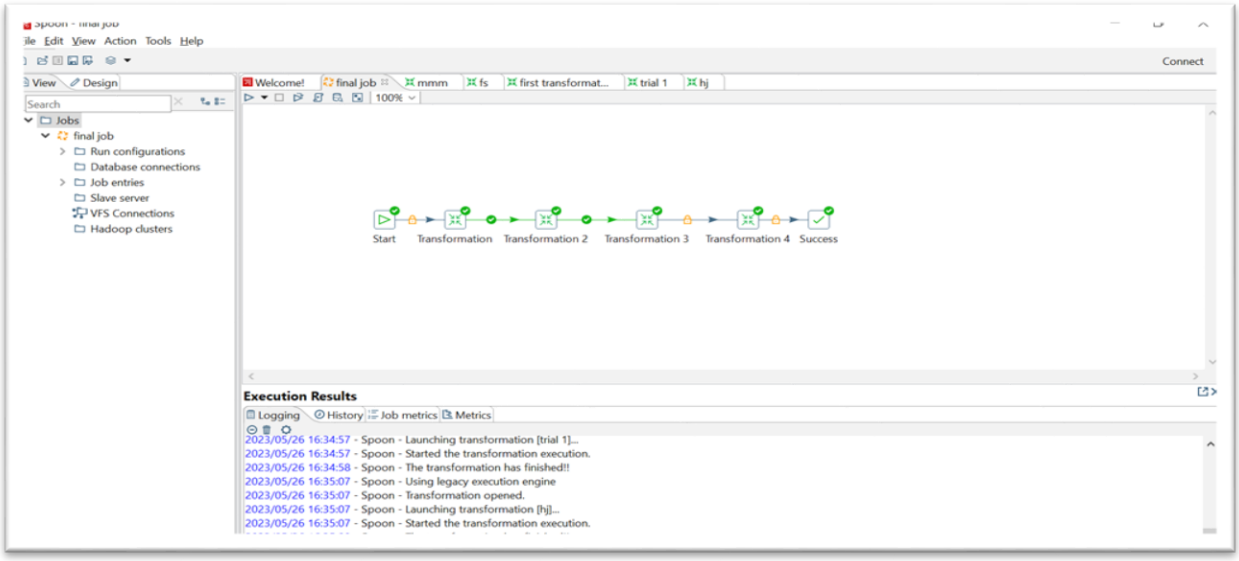
#	result	clientcode	outputfilename	fileformat	processstatus
1	"800000000000000000000"$"varshini000000000000"$"bala0000000000000000"	02	op2.csv	delimiter	ready
2	"600000000000000000000"$"sahithya000000000000"$"sanjay00000000000000"	02	op2.csv	delimiter	ready
3	"500000000000000000000"$"praveen000000000000"$"kannan00000000000000"	02	op2.csv	delimiter	ready
4	"300000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000"	02	op2.csv	delimiter	ready



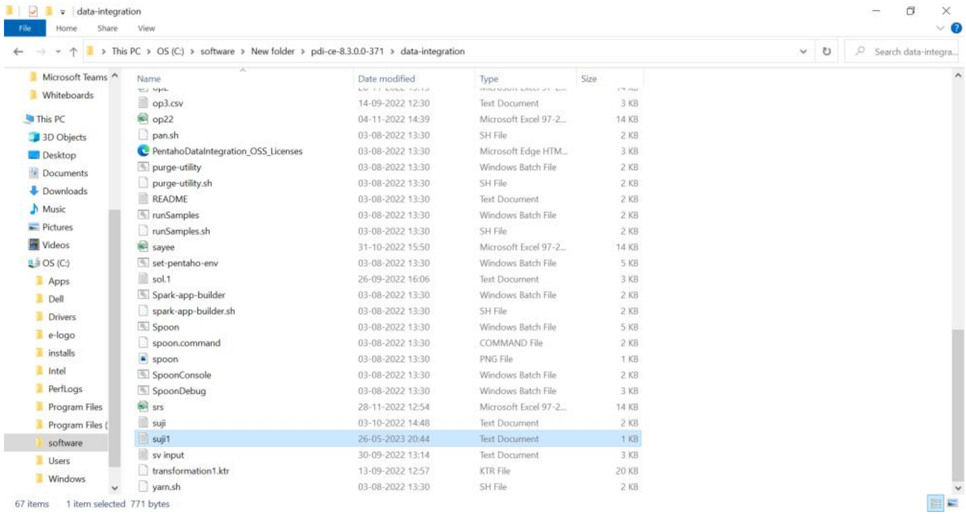
#	clientcode	outputfilename	fileformat	processstatus	result
1	02	op2.csv	delimiter	ready	800000000000000000000varshini000000000000bala0000000000000000
2	02	op2.csv	delimiter	ready	600000000000000000000sahithya000000000000sanjay00000000000000
3	02	op2.csv	delimiter	ready	500000000000000000000praveen000000000000kannan00000000000000
4	02	op2.csv	delimiter	ready	300000000000000000000kaviya00000000000000senthilkumar00000000



Now we have to run the final job transformation to get those required data in a location with our desired file format.

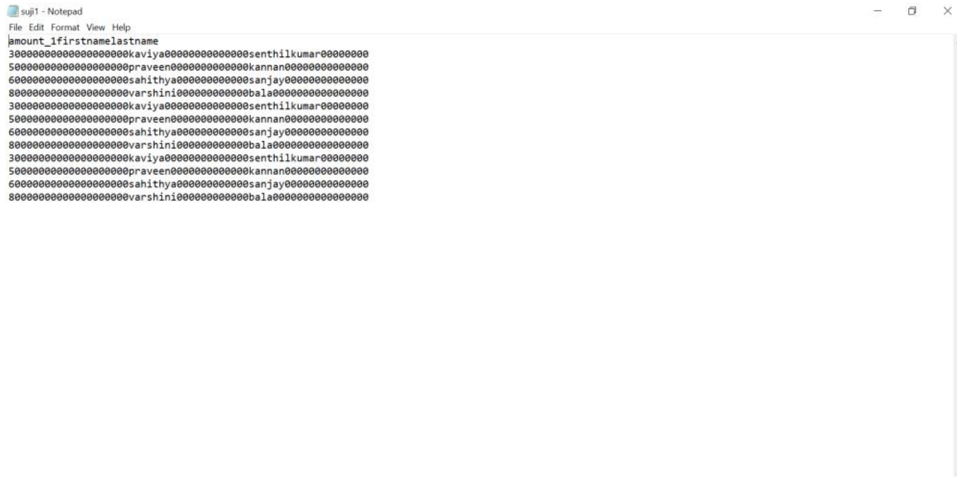


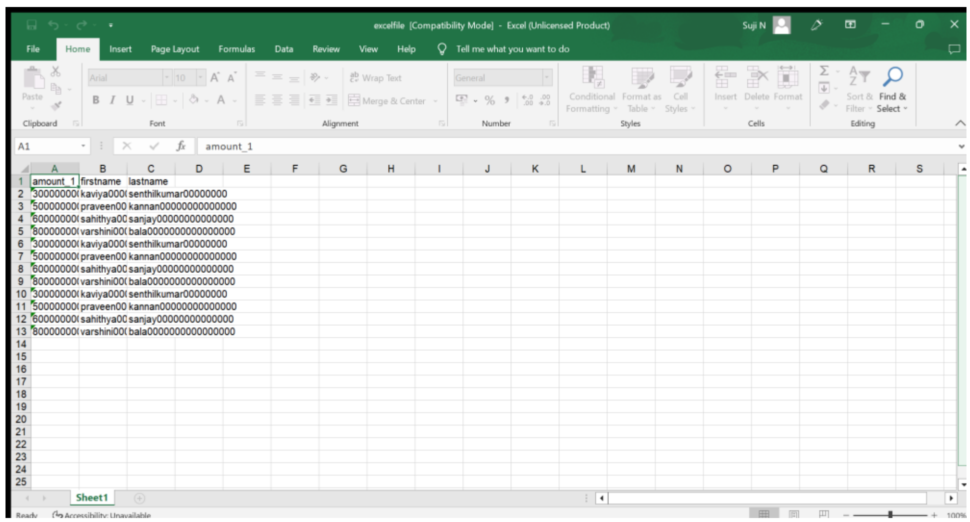
Given below are the files with the required data that generated automatically

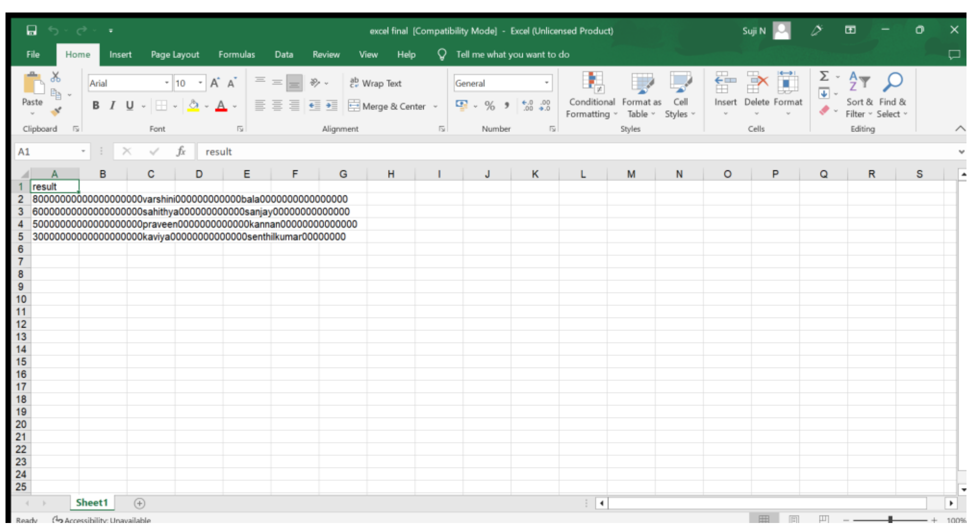

data-integration
File
Home
Share
View
This PC > OS (C:) > software > New folder > pdi-ce-8.3.0.0-371 > data-integration
Search data-integra...
Microsoft Teams
Whiteboards
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
Apps
Dell
Drivers
e-logo
installs
Intel
PerfLogs
Program Files
software
Users
Windows
Name
Date modified
Type
Size
op3.csv 14-09-2022 12:30 Text Document 3 KB
op22 04-11-2022 14:39 Microsoft Excel 97-2... 14 KB
pan.sh 03-08-2022 13:30 SH File 2 KB
PentahoDataIntegration_OSS_Licenses 03-08-2022 13:30 Microsoft Edge HTM... 3 KB
purge-utility 03-08-2022 13:30 Windows Batch File 2 KB
purge-utility.sh 03-08-2022 13:30 SH File 2 KB
README 03-08-2022 13:30 Text Document 2 KB
runSamples 03-08-2022 13:30 Windows Batch File 2 KB
runSamples.sh 03-08-2022 13:30 SH File 2 KB
sayee 31-10-2022 15:50 Microsoft Excel 97-2... 14 KB
set-pentaho-env 03-08-2022 13:30 Windows Batch File 5 KB
sol.1 26-09-2022 16:06 Text Document 3 KB
Spark-app-builder 03-08-2022 13:30 Windows Batch File 2 KB
spark-app-builder.sh 03-08-2022 13:30 SH File 2 KB
Spoon 03-08-2022 13:30 Windows Batch File 5 KB
spoon.command 03-08-2022 13:30 COMMAND File 2 KB
spoon 03-08-2022 13:30 PNG File 1 KB
SpoonConsole 03-08-2022 13:30 Windows Batch File 2 KB
SpoonDebug 03-08-2022 13:30 Windows Batch File 3 KB
srs 28-11-2022 12:54 Microsoft Excel 97-2... 14 KB
suji 03-10-2022 14:48 Text Document 2 KB
suji1 26-05-2023 20:44 Text Document 1 KB
sv input 30-09-2022 13:14 Text Document 3 KB
transformation1.ktr 13-09-2022 12:57 KTR File 20 KB
yarn.sh 03-08-2022 13:30 SH File 2 KB
67 items 1 item selected 771 bytes

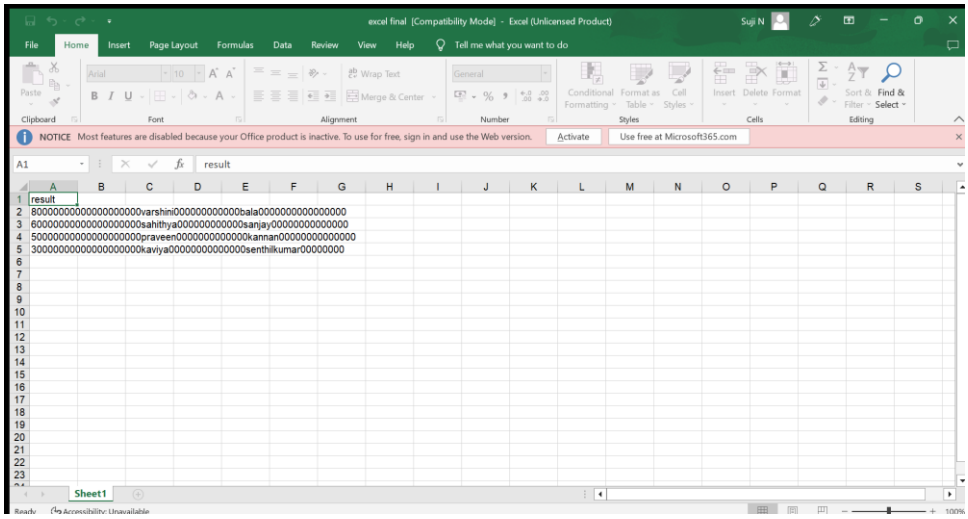

op2.csv - Notepad
File Edit Format View Help
result
"80000000000000000000"$"varshini000000000000"$"bala0000000000000000"
"60000000000000000000"$"sahithya000000000000"$"sanjay00000000000000"
"50000000000000000000"$"praveen0000000000000"$"kannan00000000000000"
"30000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000"


op2.csv - Notepad
File Edit Format View Help
amountfirstnamelastname
"30000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000"
"50000000000000000000"$"praveen0000000000000"$"kannan00000000000000"
"60000000000000000000"$"sahithya000000000000"$"sanjay00000000000000"
"80000000000000000000"$"varshini000000000000"$"bala0000000000000000"
"30000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000"
"50000000000000000000"$"praveen0000000000000"$"kannan00000000000000"
"60000000000000000000"$"sahithya000000000000"$"sanjay00000000000000"
"80000000000000000000"$"varshini000000000000"$"bala0000000000000000"
"30000000000000000000"$"kaviya00000000000000"$"senthilkumar00000000"
"50000000000000000000"$"praveen0000000000000"$"kannan00000000000000"
"60000000000000000000"$"sahithya000000000000"$"sanjay00000000000000"
"80000000000000000000"$"varshini000000000000"$"bala0000000000000000"


suji1 - Notepad
File Edit Format View Help
amount_1firstnamelastname
30000000000000000000kaviya000000000000000senthilkumar00000000
50000000000000000000praveen00000000000000kannan00000000000000
60000000000000000000sahithya0000000000000sanjay00000000000000
80000000000000000000varshini0000000000000bala0000000000000000
30000000000000000000kaviya000000000000000senthilkumar00000000
50000000000000000000praveen00000000000000kannan00000000000000
60000000000000000000sahithya0000000000000sanjay00000000000000
80000000000000000000varshini0000000000000bala0000000000000000
30000000000000000000kaviya000000000000000senthilkumar00000000
50000000000000000000praveen00000000000000kannan00000000000000
60000000000000000000sahithya0000000000000sanjay00000000000000
80000000000000000000varshini0000000000000bala0000000000000000


excelfile [Compatibility Mode] - Excel (Unlicensed Product)
A1
amount_1
amount_1 firstname lastname
Sheet1


excel final [Compatibility Mode] - Excel (Unlicensed Product)
A1
result
result
80000000000000000000varshini0000000000000bala0000000000000000
60000000000000000000sahithya0000000000000sanjay00000000000000
50000000000000000000praveen00000000000000kannan00000000000000
30000000000000000000kaviya000000000000000senthilkumar00000000
Sheet1



Finally this project is to pull data from different tables automatically according to our requirements and generate files according to our required file types with those filtered data. This works for whatever data we push into PgAdmin.