

Case study | Azure Data Bricks

Divya Sree Murali

Create a workspace and add three notebooks for each of ETL process

Notebook 1--→Extract---[Reading raw Data]

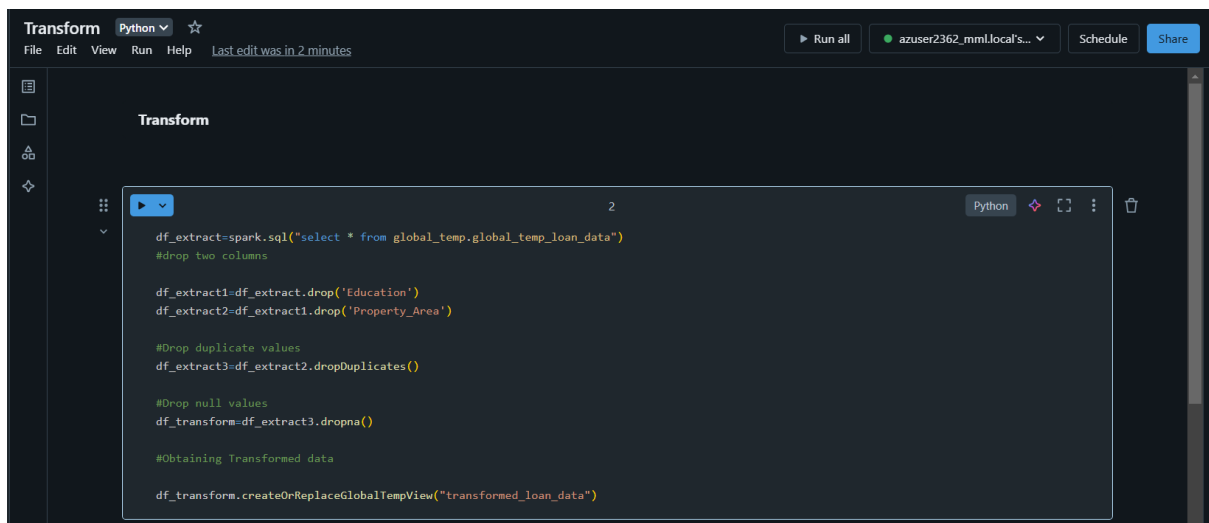


```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("Et1").getOrCreate()
filepath="/FileStore/tables/LoanData.csv"
df=spark.read.csv(filepath,header=True,inferSchema=True)
df.createOrReplaceGlobalTempView("global_temp_loan_data")
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [Loan_ID: string, Gender: string ... 11 more fields]

Notebook 2→Transform---[Apply suitable transformations on raw data]



```
df_extract=spark.sql("select * from global_temp.global_temp_loan_data")
#drop two columns

df_extract1=df_extract.drop('Education')
df_extract2=df_extract1.drop('Property_Area')

#Drop duplicate values
df_extract3=df_extract2.dropDuplicates()

#Drop null values
df_transform=df_extract3.dropna()

#Obtaining Transformed data

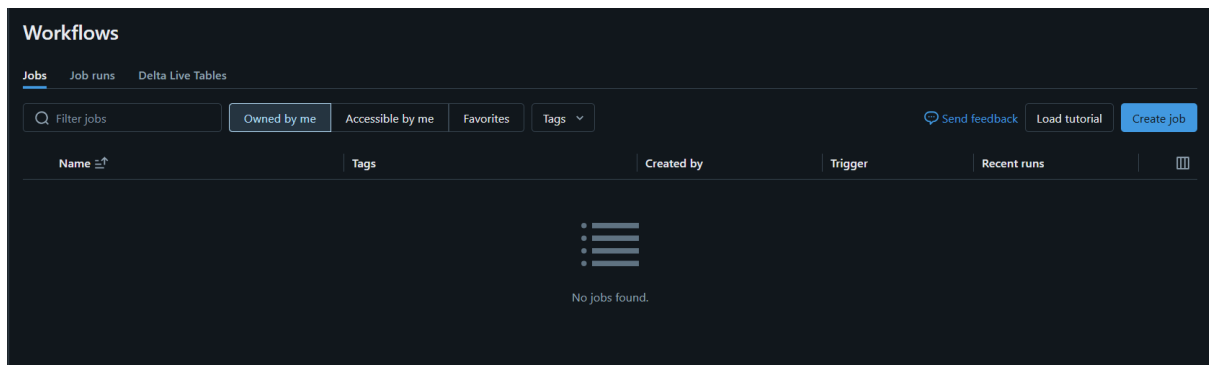
df_transform.createOrReplaceGlobalTempView("transformed_loan_data")
```

Notebook 3--→Load----[Loading the transformed data and saving as table]

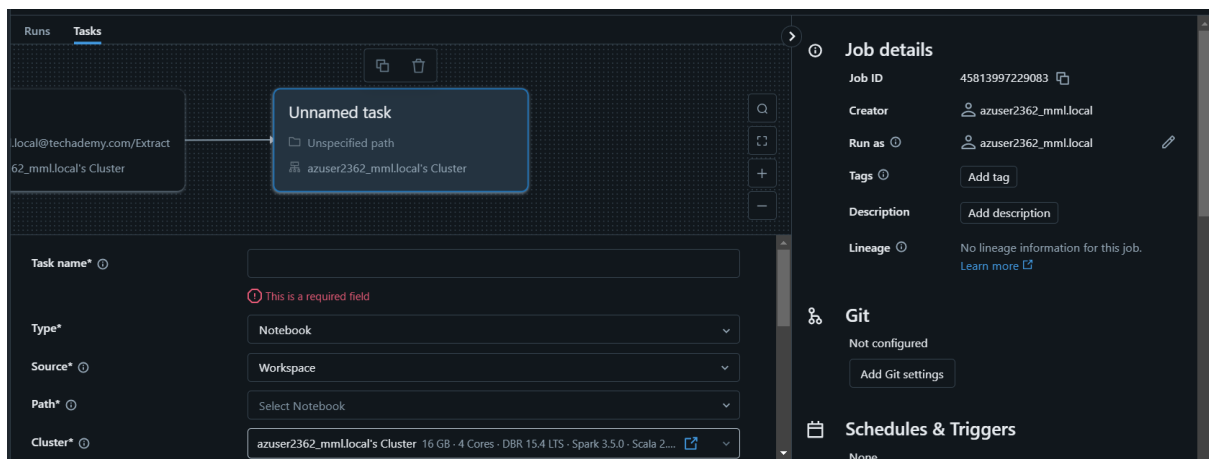


```
df_new=spark.sql("select * from global_temp.transformed_loan_data1")
df_new.write.format("delta").saveAsTable("transformed_loan_data1")
```

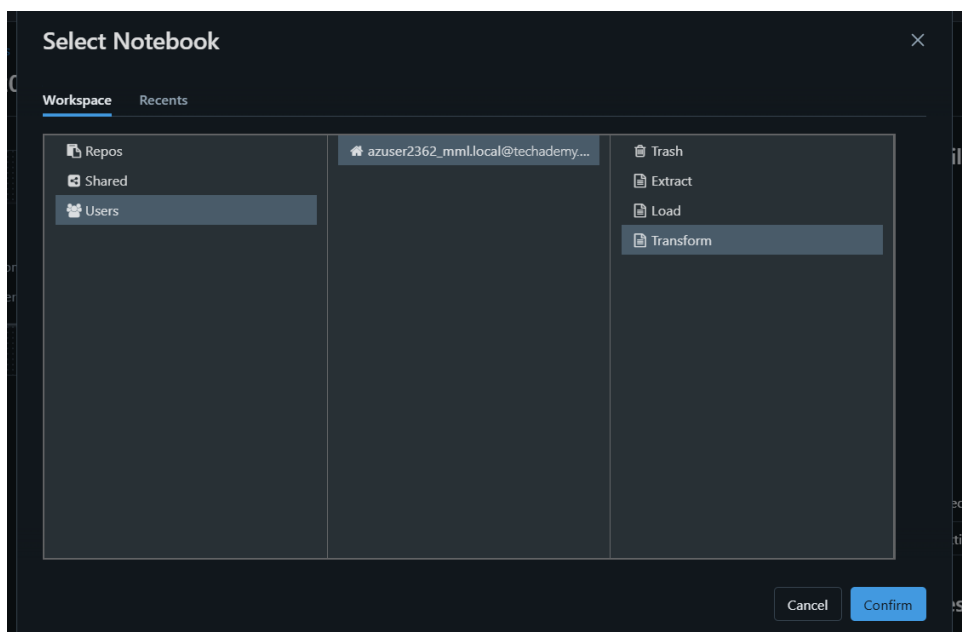
Go to workflows and select create job



First create task by adding extract notebook directory



Click on add task and add transform and load notebooks directory successively.



New Job 2024-12-10 15:54:53 ☆

Runs Tasks

Load

...362_mm1.local@techademy.com/Load

azuser2362_mm1.local's Cluster

Task name* ⓘ Load

Type* Notebook

Source* ⓘ Workspace

Path* ⓘ /Workspace/Users/azuser2362_mm1.local@techademy.com/Load

Cluster* ⓘ azuser2362_mm1.local's Cluster 16 GB · 4 Cores · DBR 15.4 LTS · Spark 3.5.0 · Scala 2....

ⓘ Jobs running on all-purpose clusters are considered all-purpose compute. [Learn more](#)

Cancel Create task

Click on run now and wait until job run gets completed.

Workflows > Jobs >

New Job 2024-12-10 15:54:53 ☆

Runs Tasks

Extract Transform Load

+ Add task

Task name* ⓘ Load

Type* Notebook

Source* ⓘ Workspace

Path* ⓘ /Workspace/Users/azuser2362_mm1.local@techademy.com/Load

Cluster* ⓘ azuser2362_mm1.local's Cluster 16 GB · 4 Cores · DBR 15.4 LTS · Spark 3.5.0 · Scala 2....

ⓘ Jobs running on all-purpose clusters are considered all-purpose compute. [Learn more](#)

Cancel Save task

Send feedback Run now

Job details

Job ID 45813997229083

Creator azuser2362_mm1.local

Run as ⓘ azuser2362_mm1.local

Tags ⓘ Add tag

Description Add description

Lineage ⓘ No lineage information for this job. [Learn more](#)

Git

Not configured

Add Git settings

Schedules & Triggers

None

Add trigger

Triggered run: 202192130978963

[View run](#)

RunsTasks

Runs

Start date

< Previous

Next >

Run total duration

0s

0s

Dec 10

Tasks

Extract

Transform

Load

Cancel runs

Start time	Run ID	Launched	Duration	Status	Error code	Run paramet...
Dec 10, 2024, 04:0...	20219213097...	Manually	27s	Runni...		

Job details

Job ID

45813997229083

Creator

azuser2362_mml.local

Run as

azuser2362_mml.local

Tags

Add tag

Description

Add description

Lineage

No lineage information for this job.
[Learn more](#)

Git

Not configured

Add Git settings

Schedules & Triggers

None

Add trigger

GraphTimeline

Extract

Running · 6s

...2_mml.local@techademy.com/Extract

azuser2362_mml.local's Cluster

→

Transform

Blocked · 0s

...ml.local@techademy.com/Transform

azuser2362_mml.local's Cluster

→

Load

Blocked · 0s

...62_mml.local@techademy.com/Load

azuser2362_mml.local's Cluster

Job run details

Job ID

45813997229083

Job run ID

367639536580122

Launched

Manually

Started

12/10/2024, 04:02:06 PM

Ended

-

Duration

9s

Queue duration

-

Status

Running - Cancel

Lineage

0 upstream tables, 1 down...

View run events

Extract

Succeeded · 18s

...2_mml.local@techademy.com/Extract

azuser2362_mml.local's Cluster

→

Transform

Running · 4s

...ml.local@techademy.com/Transform

azuser2362_mml.local's Cluster

→

Load

Blocked · 0s

...62_mml.local@techademy.com/Load

azuser2362_mml.local's Cluster

Job n

Laun

Start

Ende

Dura

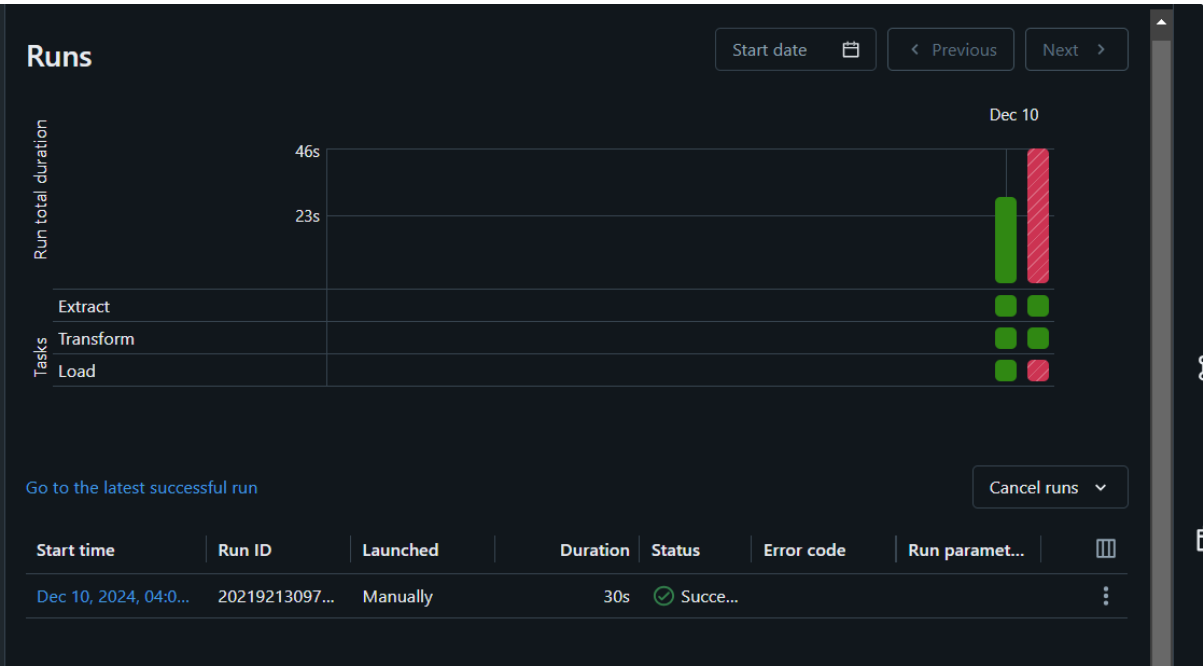
Queu

Statu

Linea

Vie

When all the notebooks run the output of the job will be stated as “succeeded”.In case if there are any error in the comands given in notebook the status appear as “Failed”



The loaded data is present in catalog as delta tables

