🧙‍♀️ Mage & 🏎️ DTC Zoomcamp

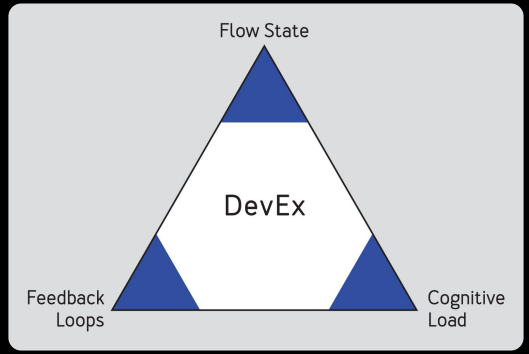
💫 Making Orchestration Magical with Mage

What is Mage?

A good orchestrator prioritizes….

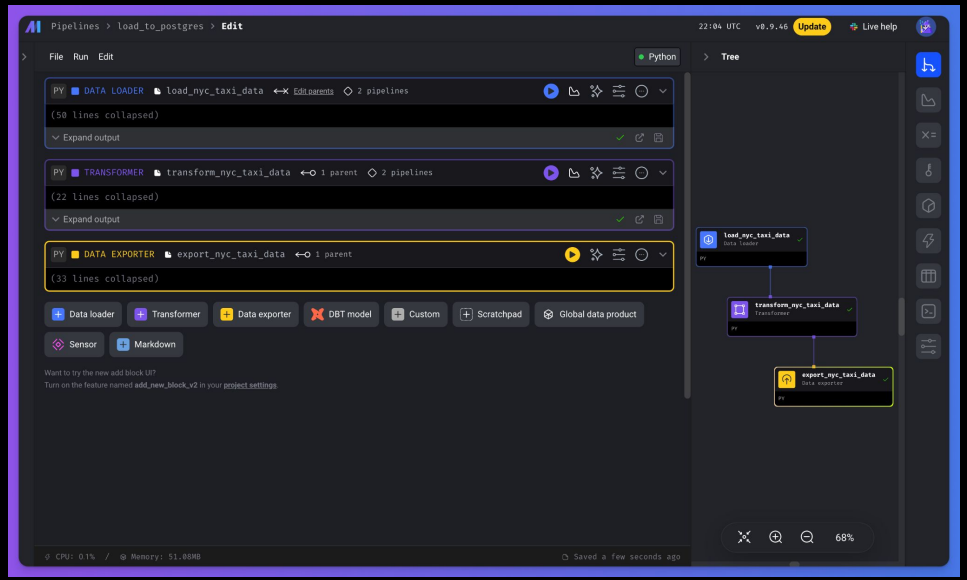
The *developer experience*

- Flow state 🌊
  - *“I need to switch between 7 tools/services.”*
- Feedback Loops 🔁
  - *“I spent 5 hours locally testing this DAG.”*
- Cognitive Load 🧱
  - How much do you need to know to do your job?



What is **Mage**?

An open-source pipeline tool for orchestrating, transforming, and integrating data 👷



👀 Sensors

🧪 Conditionals

👫 Dynamics

🪝 Webhooks

🏠 Projects

🏗️ Pipelines

🧱 Blocks

👷 Data Integration

🤝 Unified Pipelines

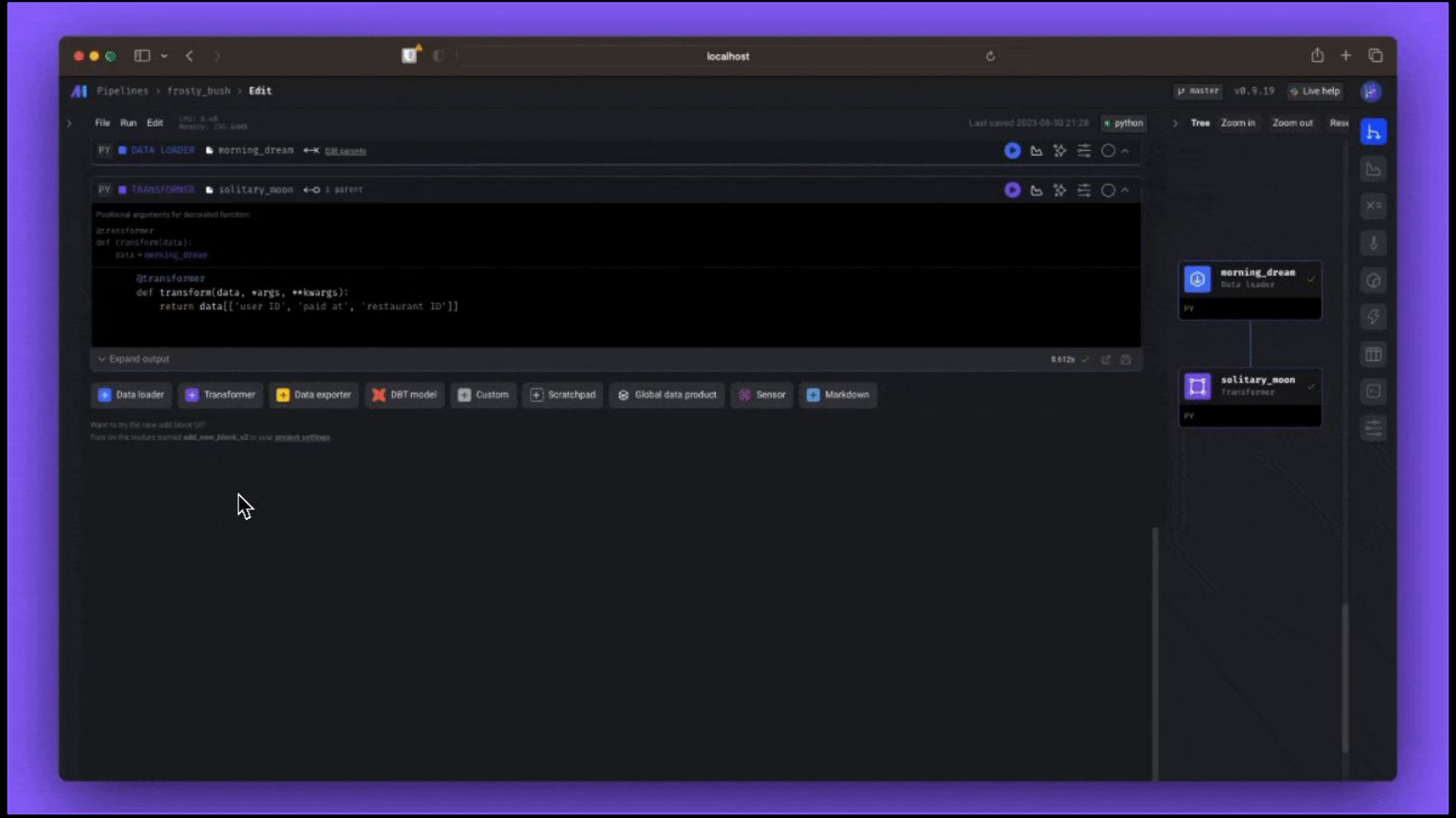
👯 Multi-user envs

✍️ Templating

🔨 Load

👷 Transform

⚙️ Export

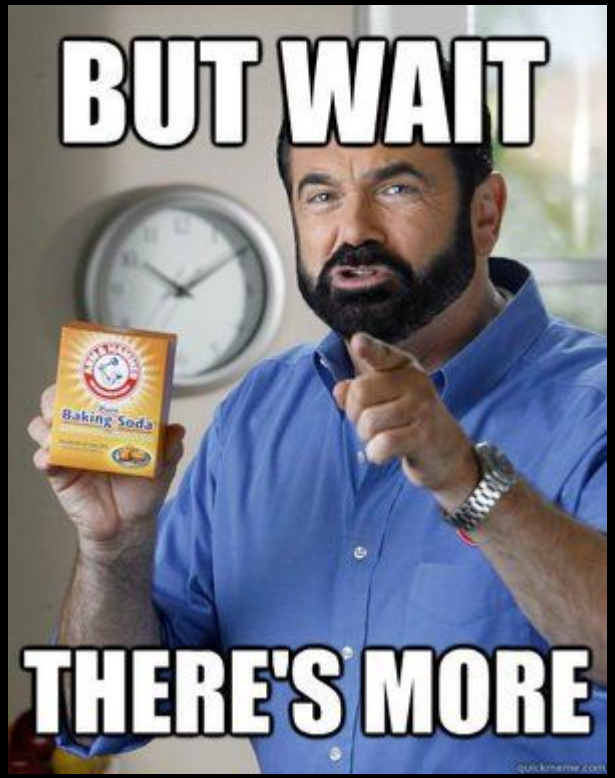

localhost
Pipelines › frosty_bush › Edit
File Run Edit
master
v0.9.19
Live help
python
Tree
Zoom in
Zoom out
DATA LOADER
morning_dream
Edit parents
TRANSFORMER
solitary_moon
1 parent
@transformer
def transform(data, *args, **kwargs):
return data[['user ID', 'paid at', 'restaurant ID']]
Expand output
Data loader
Transformer
Data exporter
DBT model
Custom
Scratchpad
Global data product
Sensor
Markdown
morning_dream
Data loader
solitary_moon
Transformer

🚀 Mage *accelerates* pipeline development

- Hybrid environment
  - Use our GUI for interactive development (or don't, I like VSCode)
  - Use blocks as testable, reusable pieces of code.
- Improved DevEx
  - Code and test in parallel.
  - Reduce your dependencies, switch tools less, be efficient.

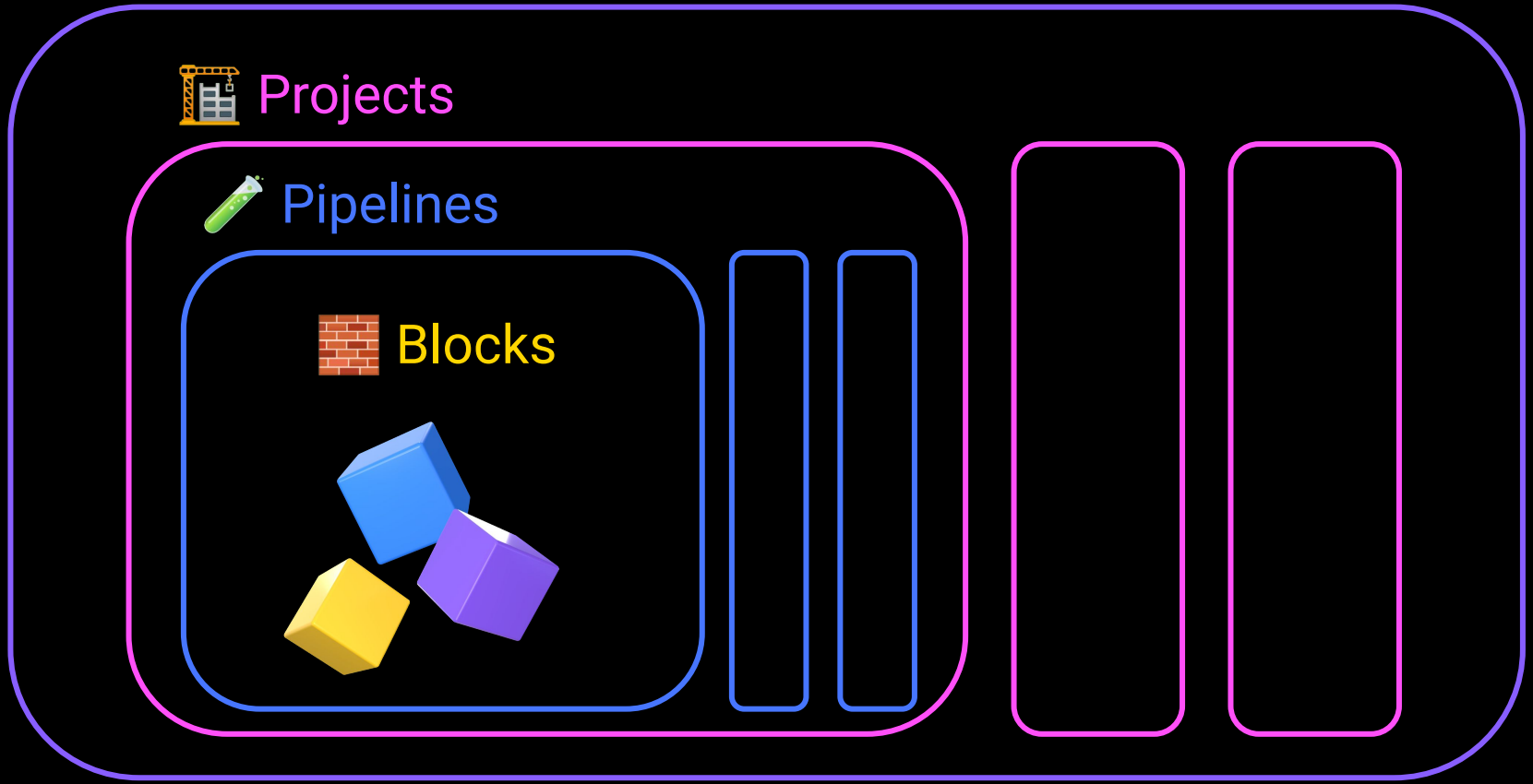
👨‍💻 Engineering best-practices built-in

- 🧪 In-line testing and debugging
  - Familiar, notebook-style format
- 🔎 Fully-featured observability
  - Transformation *in one place*: dbt models, streaming, *& more*.
- 🏜️ DRY principles
  - No more 🍝 DAGs with duplicate functions and weird imports
  - DEaaS (sorry, I had to 😅)


BUT WAIT
Baking Soda
THERE'S MORE

Reduce time in undifferentiated work.

MAGE



🏗️ Projects

- A project forms the basis for all the work you can do in Mage—you can think of it like a GitHub repo.
- It contains the code for all of your pipelines, blocks, and other assets.
- A Mage instance has one or more projects

🧪 Pipelines

- A pipeline is a workflow that executes some data operation—maybe extracting, transforming, and loading data from an API. They’re also called DAGs on other platforms
- In Mage, pipelines can contain *Blocks* (written in SQL, Python, or R) and charts.
- Each pipeline is represented by a YAML file in the “pipelines” folder of your project.

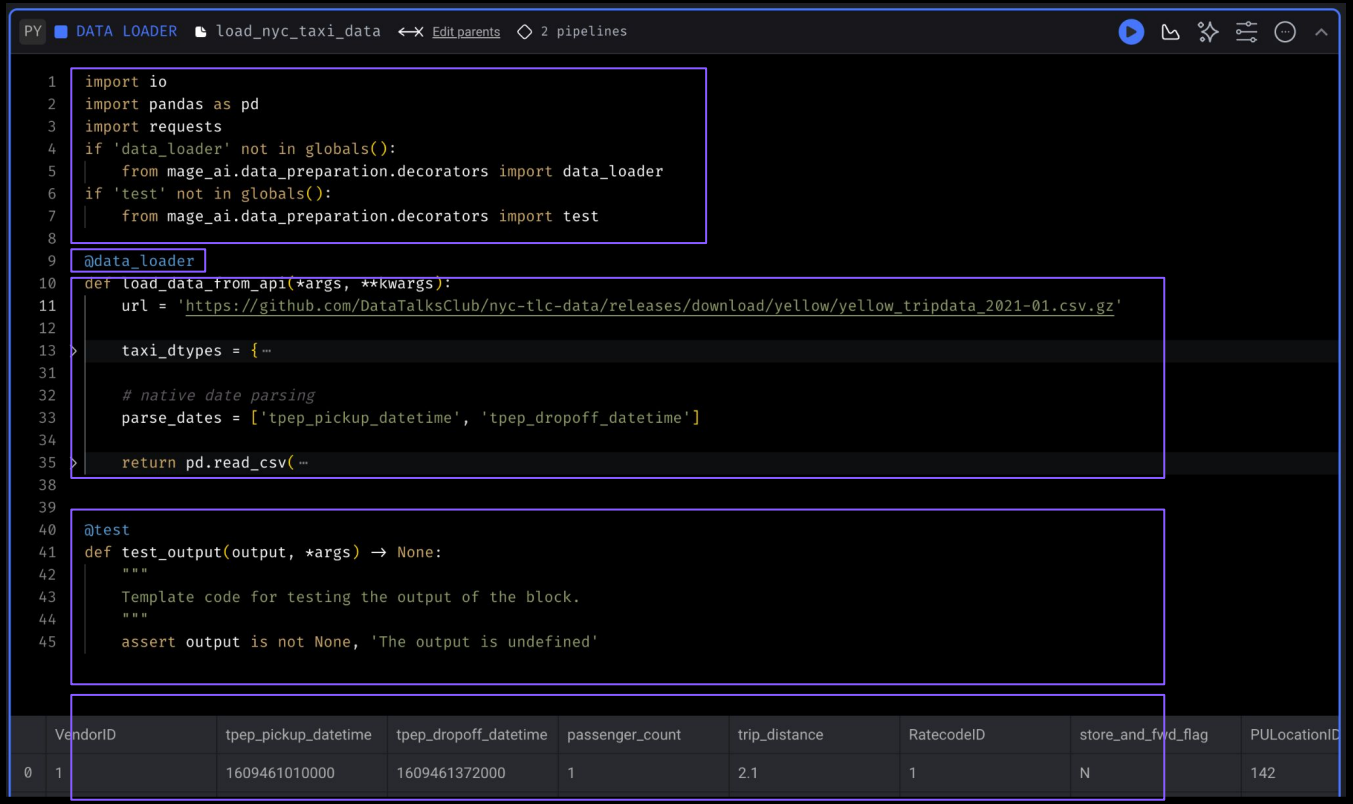
Blocks

- A block is a file that can be executed independently or within a pipeline.
- Together, blocks form Directed Acyclic Graphs (DAGs), which we call pipelines.
- A block won’t start running in a pipeline until all its upstream dependencies are met.

Blocks continued

- Blocks are reusable, atomic pieces of code that perform certain actions.
- Changing one block will change it everywhere it’s used, but **don’t worry**, it’s easy to detach blocks to separate instances if necessary.
- Blocks can be used to perform a variety of actions, from simple data transformations to complex machine learning models.

Anatomy of a Block



Imports

Decorator

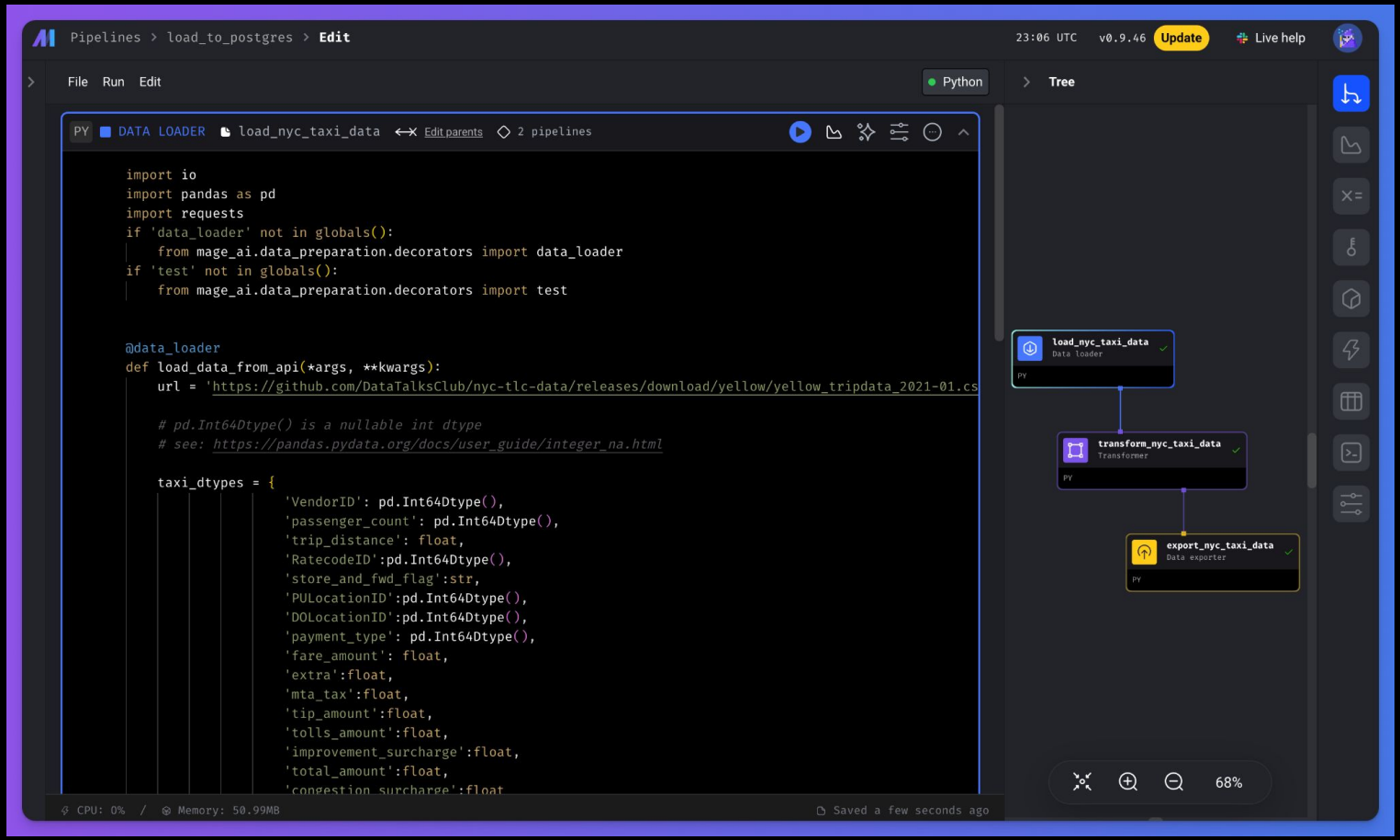
Function*

Assertion

DataFrame

*returns df

Up next: Configuring Mage & running a pipeline


Pipelines > load_to_postgres > Edit
23:06 UTC
v0.9.46
Update
Live help
File
Run
Edit
Python
Tree
PY
DATA LOADER
load_nyc_taxi_data
Edit parents
2 pipelines
import io
import pandas as pd
import requests
if 'data_loader' not in globals():
from mage_ai.data_preparation.decorators import data_loader
if 'test' not in globals():
from mage_ai.data_preparation.decorators import test
@data_loader
def load_data_from_api(*args, **kwargs):
url = 'https://github.com/DataTalksClub/nyc-tlc-data/releases/download/yellow/yellow_tripdata_2021-01.cs
# pd.Int64Dtype() is a nullable int dtype
# see: https://pandas.pydata.org/docs/user_guide/integer_na.html
taxi_dtypes = {
'VendorID': pd.Int64Dtype(),
'passenger_count': pd.Int64Dtype(),
'trip_distance': float,
'RatecodeID':pd.Int64Dtype(),
'store_and_fwd_flag':str,
'PULocationID':pd.Int64Dtype(),
'DOLocationID':pd.Int64Dtype(),
'payment_type': pd.Int64Dtype(),
'fare_amount': float,
'extra':float,
'mta_tax':float,
'tip_amount':float,
'tolls_amount':float,
'improvement_surcharge':float,
'total_amount':float,
'congestion_surcharge':float
load_nyc_taxi_data
Data loader
transform_nyc_taxi_data
Transformer
export_nyc_taxi_data
Data exporter
68%
CPU: 0% / Memory: 50.99MB
Saved a few seconds ago