

```
In [1]: pip install dask modin[ray] pandas pyyaml
```

```
Requirement already satisfied: dask in /opt/anaconda3/lib/python3.11/site-packages (2023.11.0)
Requirement already satisfied: pandas in /opt/anaconda3/lib/python3.11/site-packages (2.2.2)
Requirement already satisfied: pyyaml in /opt/anaconda3/lib/python3.11/site-packages (6.0.1)
Requirement already satisfied: modin[ray] in /opt/anaconda3/lib/python3.11/site-packages (0.31.0)
Requirement already satisfied: click>=8.1 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (8.1.7)
Requirement already satisfied: cloudpickle>=1.5.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (2.2.1)
Requirement already satisfied: fsspec>=2021.09.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (2023.10.0)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (23.1)
Requirement already satisfied: partd>=1.2.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (1.4.1)
Requirement already satisfied: toolz>=0.10.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (0.12.0)
Requirement already satisfied: importlib-metadata>=4.13.0 in /opt/anaconda3/lib/python3.11/site-packages (from dask) (7.0.1)
Requirement already satisfied: numpy>=1.22.4 in /opt/anaconda3/lib/python3.11/site-packages (from modin[ray]) (1.26.4)
Requirement already satisfied: psutil>=5.8.0 in /opt/anaconda3/lib/python3.11/site-packages (from modin[ray]) (5.9.0)
Requirement already satisfied: ray!=2.5.0,>=2.1.0 in /opt/anaconda3/lib/python3.11/site-packages (from modin[ray]) (2.33.0)
Requirement already satisfied: pyarrow>=10.0.1 in /opt/anaconda3/lib/python3.11/site-packages (from modin[ray]) (14.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/anaconda3/lib/python3.11/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/anaconda3/lib/python3.11/site-packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in /opt/anaconda3/lib/python3.11/site-packages (from pandas) (2023.3)
Requirement already satisfied: zipp>=0.5 in /opt/anaconda3/lib/python3.11/site-packages (from importlib-metadata>=4.13.0->dask) (3.17.0)
Requirement already satisfied: locket in /opt/anaconda3/lib/python3.11/site-packages (from partd>=1.2.0->dask) (1.0.0)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.11/site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Requirement already satisfied: filelock in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (3.13.1)
Requirement already satisfied: jsonschema in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (4.19.2)
Requirement already satisfied: msgpack<2.0.0,>=1.0.0 in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (1.0.3)
Requirement already satisfied: protobuf!=3.19.5,>=3.15.3 in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (3.20.3)
Requirement already satisfied: aiosignal in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (1.2.0)
Requirement already satisfied: frozenlist in /opt/anaconda3/lib/python3.11
```

1/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (1.4.0)
Requirement already satisfied: requests in /opt/anaconda3/lib/python3.11/site-packages (from ray!=2.5.0,>=2.1.0->modin[ray]) (2.31.0)
Requirement already satisfied: attrs>=22.2.0 in /opt/anaconda3/lib/python3.11/site-packages (from jsonschema->ray!=2.5.0,>=2.1.0->modin[ray]) (23.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /opt/anaconda3/lib/python3.11/site-packages (from jsonschema->ray!=2.5.0,>=2.1.0->modin[ray]) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in /opt/anaconda3/lib/python3.11/site-packages (from jsonschema->ray!=2.5.0,>=2.1.0->modin[ray]) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in /opt/anaconda3/lib/python3.11/site-packages (from jsonschema->ray!=2.5.0,>=2.1.0->modin[ray]) (0.10.6)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/anaconda3/lib/python3.11/site-packages (from requests->ray!=2.5.0,>=2.1.0->modin[ray]) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/python3.11/site-packages (from requests->ray!=2.5.0,>=2.1.0->modin[ray]) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/anaconda3/lib/python3.11/site-packages (from requests->ray!=2.5.0,>=2.1.0->modin[ray]) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /opt/anaconda3/lib/python3.11/site-packages (from requests->ray!=2.5.0,>=2.1.0->modin[ray]) (2024.6.2)
Note: you may need to restart the kernel to use updated packages.

In [2]: `pip install -U ipywidgets`

Requirement already satisfied: ipywidgets in /opt/anaconda3/lib/python3.11/site-packages (8.1.3)

Requirement already satisfied: comm>=0.1.3 in /opt/anaconda3/lib/python3.11/site-packages (from ipywidgets) (0.2.2)

Requirement already satisfied: ipython>=6.1.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipywidgets) (8.20.0)

Requirement already satisfied: traitlets>=4.3.1 in /opt/anaconda3/lib/python3.11/site-packages (from ipywidgets) (5.7.1)

Requirement already satisfied: widgetsnbextension~=4.0.11 in /opt/anaconda3/lib/python3.11/site-packages (from ipywidgets) (4.0.11)

Requirement already satisfied: jupyterlab-widgets~=3.0.11 in /opt/anaconda3/lib/python3.11/site-packages (from ipywidgets) (3.0.11)

Requirement already satisfied: decorator in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (5.1.1)

Requirement already satisfied: jedi>=0.16 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (0.18.1)

Requirement already satisfied: matplotlib-inline in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (0.1.6)

Requirement already satisfied: prompt-toolkit<3.1.0,>=3.0.41 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (3.0.43)

Requirement already satisfied: pygments>=2.4.0 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (2.15.1)

Requirement already satisfied: stack-data in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (0.2.0)

Requirement already satisfied: pexpect>4.3 in /opt/anaconda3/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (4.8.0)

Requirement already satisfied: parso<0.9.0,>=0.8.0 in /opt/anaconda3/lib/python3.11/site-packages (from jedi>=0.16->ipython>=6.1.0->ipywidgets) (0.8.3)

Requirement already satisfied: ptyprocess>=0.5 in /opt/anaconda3/lib/python3.11/site-packages (from pexpect>4.3->ipython>=6.1.0->ipywidgets) (0.7.0)

Requirement already satisfied: wcwidth in /opt/anaconda3/lib/python3.11/site-packages (from prompt-toolkit<3.1.0,>=3.0.41->ipython>=6.1.0->ipywidgets) (0.2.5)

Requirement already satisfied: executing in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=6.1.0->ipywidgets) (0.8.3)

Requirement already satisfied: asttokens in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=6.1.0->ipywidgets) (2.0.5)

Requirement already satisfied: pure-eval in /opt/anaconda3/lib/python3.11/site-packages (from stack-data->ipython>=6.1.0->ipywidgets) (0.2.2)

Requirement already satisfied: six in /opt/anaconda3/lib/python3.11/site-packages (from asttokens->stack-data->ipython>=6.1.0->ipywidgets) (1.16.0)

Note: you may need to restart the kernel to use updated packages.

Reading the file:

```
In [4]: import pandas as pd
import time

start_time = time.time()
df_pandas = pd.read_csv('Train.csv')
print(f"Pandas: {time.time() - start_time} seconds")
```

```
/var/folders/64/c_kln_751177rl4_r01mrtzm0000gn/T/ipykernel_31551/352804403
0.py:5: DtypeWarning: Columns (5,8,11,15) have mixed types. Specify dtype
option on import or set low_memory=False.
```

```
df_pandas = pd.read_csv('Train.csv')
```

```
Pandas: 59.813443183898926 seconds
```

```
In [5]: import modin.pandas as mpd
import ray
```

```
ray.init()
```

```
start_time = time.time()
```

```
df_modin = mpd.read_csv('Train.csv')
```

```
print(f"Modin: {time.time() - start_time} seconds")
```

```
2024-07-30 06:55:33,842 INFO worker.py:1781 -- Started a local Ray instanc
e.
```

```
Modin: 40.09968304634094 seconds
```

```
(raylet) Spilled 2442 MiB, 13 objects, write throughput 463 MiB/s. Set RAY
_verbose_spill_logs=0 to disable this message.
```

```
UserWarning: `read_*` implementation has mismatches with pandas:
```

```
Data types of partitions are different! Please refer to the troubleshootin
g section of the Modin documentation to fix this issue.
```

```
In [6]: import yaml
```

```
schema = {
```

```
    'separator': ',',
```

```
    'columns': df_pandas.columns.tolist()
```

```
}
```

```
with open('schema.yaml', 'w') as file:
```

```
    yaml.dump(schema, file)
```

```
In [7]: # Display the column names to check for exact matches
```

```
print(df_pandas.columns)
```

```
Index(['fecha_dato', 'ncodpers', 'ind_empleado', 'pais_residencia', 'sex
o',
```

```
      'age', 'fecha_alta', 'ind_nuevo', 'antiguedad', 'indrel',
```

```
      'ult_fec_cli_1t', 'indrel_1mes', 'tiprel_1mes', 'indresi', 'index
```

```
t',
```

```
      'conyuemp', 'canal_entrada', 'indfall', 'tipodom', 'cod_prov',
```

```
      'nomprov', 'ind_actividad_cliente', 'renta', 'segmento',
```

```
      'ind_ahor_fin_ult1', 'ind_aval_fin_ult1', 'ind_cco_fin_ult1',
```

```
      'ind_cder_fin_ult1', 'ind_cno_fin_ult1', 'ind_ctju_fin_ult1',
```

```
      'ind_ctma_fin_ult1', 'ind_ctop_fin_ult1', 'ind_ctpp_fin_ult1',
```

```
      'ind_deco_fin_ult1', 'ind_deme_fin_ult1', 'ind_dela_fin_ult1',
```

```
      'ind_ecue_fin_ult1', 'ind_fond_fin_ult1', 'ind_hip_fin_ult1',
```

```
      'ind_plan_fin_ult1', 'ind_pres_fin_ult1', 'ind_reca_fin_ult1',
```

```
      'ind_tjcr_fin_ult1', 'ind_valo_fin_ult1', 'ind_viv_fin_ult1',
```

```
      'ind_nomina_ult1', 'ind_nom_pens_ult1', 'ind_recibo_ult1'],
```

```
dtype='object')
```

```
In [8]: # Display the updated DataFrame
```

```
print(df_pandas.head())
```

	fecha_datos	ncodpers	ind_employed	pais_residencia	sexo	age	fecha_alta
0	2015-01-28	1375586	N	ES	H	35	2015-01-12
1	2015-01-28	1050611	N	ES	V	23	2012-08-10
2	2015-01-28	1050612	N	ES	V	23	2012-08-10
3	2015-01-28	1050613	N	ES	H	22	2012-08-10
4	2015-01-28	1050614	N	ES	V	23	2012-08-10

	ind_nuevo	antiguedad	indrel	...	ind_hip_fin_ult1	ind_plan_fin_ult1	\
0	0.0	6	1.0	...	0	0	
1	0.0	35	1.0	...	0	0	
2	0.0	35	1.0	...	0	0	
3	0.0	35	1.0	...	0	0	
4	0.0	35	1.0	...	0	0	

	ind_pres_fin_ult1	ind_reca_fin_ult1	ind_tjcr_fin_ult1	ind_valo_fin_ult1
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

	ind_viv_fin_ult1	ind_nomina_ult1	ind_nom_pens_ult1	ind_recibo_ult1
0	0	0.0	0.0	0
1	0	0.0	0.0	0
2	0	0.0	0.0	0
3	0	0.0	0.0	0
4	0	0.0	0.0	0

[5 rows x 48 columns]

In []: