1.

import pandas as pd

df = pd.read\_table(‘/Users/jinyu/Downloads/drive-download-20220810T202736Z-001/people/people\_1.txt’, sep = ‘\t’)

df1= df.strip()

df2 = df1(‘Address’).apply(lambda x : x.lstrip(‘No.#’))

df3= df2(‘Phone’).apply(lambda x: x.replace(“-”, “”))

df4= df3(’FirstName’).apply(lambda x: x.upper())

df5= df4(’LastName’).apply(lambda x: x.upper())

dataframe\_p1= df5.drop\_duplicates()

DF = pd.read\_table(‘/Users/jinyu/Downloads/drive-download-20220810T202736Z-001/people/people\_2.txt’, sep = ‘\t’)

DF1= DF.strip()

DF2 = DF1(‘Address’).apply(lambda x : x.lstrip(‘No.#’))

DF3= DF2(‘Phone’).apply(lambda x: x.replace(“-”, “”))

DF4= DF3(’FirstName’).apply(lambda x: x.upper())

DF5= DF4(’LastName’).apply(lambda x: x.upper())

dataframe\_p2= DF5.drop\_duplicates()

2.

import pandas as pd

number\_of\_splits=7

for i in range(0,number\_of\_splits+1):

word =i+1

print(word)

json\_file =f”movie.json"

csv\_file =f”movie.csv"

df = pd.read\_json (fr’/Users/jinyu/Downloads/drive-download-20220810T202736Z-001/{json\_file}')

df.to\_csv (fr'/Users/jinyu/Downloads/drive-download-20220810T202736Z-001/{csv\_file}', index = None)

3.

Cloud computing:

Software:

Applications and data: data, apps

Platform:

Runtime: IIS, Docker

Middleware: Software

Operating System: Windows, Linux

Infrastructure:

Virtualization: VM

Servers: Memory, Motherboard, CPU

Networking: routers, switches, internet

Storage: HDD, SSD

On-Premises:

Cloud provider manages nothing.

You manage everything:

Infrastructure

Platform

Software

IaaS:

Cloud provider manages infrastructure

You manage platform and software

Use cases:

Migration of workloads

Test and development

Storage, backups and recovery

Virtual Machine, Virtual Network, Managed Disk

PaaS:

Cloud provider manages infrastructure and platform

You manage software.

Use cases:

Development framework

Analytics and business intelligence

SQL, App Service, Logic Apps, Function Apps

SaaS:

Cloud provider manages everything

You manage nothing

Use cases:

Buying of the shell apps

One Drive, Outlook, Skype

4.

ETL: Extract, Transform, Load

ELT: Extract, Load, Transform

The difference between ETL and ELT is ETL will transforms data on a separate processing server, while ELT does that within the data warehouse itself.

And ETL does not transfer raw data into the data warehouse, however, ELT sends raw data directly to the data warehouse.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ETL | ELT |  |
| PROS | A time intensive process | Faster |  |
|  | Small data sets with complicated transformation requirements | Large datasets that require speed and efficiency |  |
|  | Has existed for over 20 years; its practices and protocols are well-known and documented | Raw data can be required endlessly |  |
|  | Pre-load transformation ca eliminate PII | Simplified data stack costs less |  |
|  |  | The maintenance burden is reduced |  |
| CONS | Does not have data lake compatibility | Does have data lake compatibility |  |
|  | Large datasets that require speed and efficiency | Small data sets with complicated transformation requirements |  |
|  | Raw data cannot be requeried | A newer form of data integration; less documentation and experience |  |
|  | Separate servers can create cost issues | Direct loading of data required more privacy safeguards |  |
|  | Secondary processing server adds to the maintenance burden |  |  |