etl

June 2, 2025

1 PART I

[19]:
```python
import pandas as pd

# Load the dataset from the mounted volume
file_path = 'healthcare_dataset-20250506.csv'
df = pd.read_csv(file_path, sep=';')

# 1) Show the overall shape (rows × columns)
print("Shape of DataFrame:", df.shape)

# 2) List each column name with its data type
print("\nColumn Names and Data Types:")
print(df.dtypes)

# 3) Display the first 5 rows as a quick sample
print("\nFirst 5 rows of the dataset:")
print(df.head())
```

```
Shape of DataFrame: (55500, 15)

Column Names and Data Types:
Name                   object
Age                     int64
Gender                 object
Blood Type             object
Medical Condition      object
Date of Admission      object
Doctor                 object
Hospital               object
Insurance Provider     object
Billing Amount        float64
Room Number             int64
Admission Type         object
Discharge Date         object
Medication             object
Test Results           object
dtype: object
```

```
First 5 rows of the dataset:
            Name  Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn   30    Male         B-            Cancer        31/01/2024
1   LesLie TErRy   62    Male         A+           Obesity        20/08/2019
2    DaNnY sMitH   76  Female         A-           Obesity        22/09/2022
3   andrEw waTtS   28  Female         O+          Diabetes        18/11/2020
4  adrIENNE bEll   43  Female        AB+            Cancer        19/09/2022

             Doctor                     Hospital Insurance Provider  \
0     Matthew Smith             Sons and Miller         Blue Cross
1   Samantha Davies                      Kim Inc           Medicare
2  Tiffany Mitchell                     Cook PLC              Aetna
3       Kevin Wells  Hernandez Rogers and Vang,           Medicare
4    Kathleen Hanna                 White-White              Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0    18856.281306          328         Urgent     02/02/2024  Paracetamol
1    33643.327287          265      Emergency     26/08/2019    Ibuprofen
2    27955.096079          205      Emergency     07/10/2022      Aspirin
3    37909.782410          450       Elective     18/12/2020    Ibuprofen
4    14238.317814          458         Urgent     09/10/2022   Penicillin

   Test Results
0        Normal
1  Inconclusive
2        Normal
3      Abnormal
4      Abnormal
```

2 PART II

[2]:
```
!pip install pymongo
```

```
Collecting pymongo
  Downloading pymongo-4.13.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (22 kB)
Collecting dnspython<3.0.0,>=1.16.0 (from pymongo)
  Downloading dnspython-2.7.0-py3-none-any.whl.metadata (5.8 kB)
Downloading
pymongo-4.13.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.4
MB)
                         1.4/1.4 MB
2.6 MB/s eta 0:00:00:00:0100:01
Downloading dnspython-2.7.0-py3-none-any.whl (313 kB)
                         313.6/313.6 kB
1.0 MB/s eta 0:00:00:00:01
```

```
Installing collected packages: dnspython, pymongo
Successfully installed dnspython-2.7.0 pymongo-4.13.0
```

[20]:
```python
from pymongo import MongoClient
import os
from datetime import datetime

MONGO_USER = os.getenv("MONGO_INITDB_ROOT_USERNAME", "root")
MONGO_PASS = os.getenv("MONGO_INITDB_ROOT_PASSWORD", "password")
MONGO_HOST = os.getenv("MONGO_HOST", "mongo")
MONGO_DB   = "healthcare"

mongo_uri = f"mongodb://{MONGO_USER}:{MONGO_PASS}@{MONGO_HOST}:27017/"
client = MongoClient(mongo_uri)
db = client[MONGO_DB]
count = db.patients.count_documents({})
print("Total patients in collection:", count)
```

```
Total patients in collection: 55500
```

[27]:
```python
from datetime import datetime

cutoff = datetime(2023, 1, 1)

cursor = db.patients.find(
    {"date_of_admission": {"$gt": cutoff}}
).limit(20)

for doc in cursor:
    name = doc.get("name", "")
    doa = doc.get("date_of_admission")
    print(f"{name:30s} admitted {doa.date()}")
```

```
Bobby JacksOn                  admitted 2024-01-31
EMILY JOHNSOn                  admitted 2023-12-20
aaRon MARtiNeZ                 admitted 2023-08-13
tIMOTHY burNs                  admitted 2023-06-28
cathy sMaLl                    admitted 2023-12-23
jOSHUA OLiVer                  admitted 2023-10-03
WilLIAM cOOPEr                 admitted 2023-05-18
Erin oRTEga                    admitted 2023-05-24
kyLE bEnneTT                   admitted 2023-09-09
mIchael LiU                    admitted 2024-04-05
TAmARa hErNAndez               admitted 2023-08-17
mR. DAVID pIERce Md            admitted 2023-11-05
beThaNY MoOrE                  admitted 2023-04-09
Kim ScOtt                      admitted 2024-04-07
jOhN hARTmAN                   admitted 2023-01-07
```

```
MicHAEl MillEr                admitted 2024-02-06
kEVIn SiMmoNs Jr.             admitted 2023-12-28
JONathAn yaTeS                admitted 2023-07-24
AdriaN BuckLEY                admitted 2023-10-11
tiMOThY myers                 admitted 2024-03-02
```

[22]:
```
count_over_50 = db.patients.count_documents({"age": {"$gt": 50}})
print("Patients older than 50:", count_over_50)
```

```
Patients older than 50: 28667
```

[23]:
```
thomas_count = db.patients.count_documents({
    "name": {"$regex": r"^Thomas\s", "$options": "i"}
})
print("Patients with first name 'Thomas':", thomas_count)
```

```
Patients with first name 'Thomas': 397
```

[24]:
```
pipeline = [
    {"$group": {"_id": "$medical_condition", "count": {"$sum": 1}}},
    {"$sort": {"_id": 1}}
]
results = list(db.patients.aggregate(pipeline))

print("Count per distinct medical condition:")
for r in results:
    print(f"  {r['_id']:15s}: {r['count']}")
```

```
Count per distinct medical condition:
  Arthritis      : 9308
  Asthma         : 9185
  Cancer         : 9227
  Diabetes       : 9304
  Hypertension   : 9245
  Obesity        : 9231
```

[25]:
```
pipeline = [
    {"$group": {"_id": "$medication", "count": {"$sum": 1}}},
    {"$sort": {"_id": 1}}
]
results = list(db.patients.aggregate(pipeline))

print("Medication usage counts:")
for r in results:
    print(f"  {r['_id']:12s}: {r['count']}")
```

```
Medication usage counts:
  Aspirin     : 11094
```

```
  Ibuprofen   : 11127
  Lipitor     : 11140
  Paracetamol : 11071
  Penicillin  : 11068
```

[26]:
```python
lipitor_cursor = db.patients.find({"medication": "Lipitor"})

lipitor_list = list(lipitor_cursor)
print(f"Total Lipitor patients: {len(lipitor_list)}\n")

print("Sample Lipitor patients (first 5):")
for doc in lipitor_list[:5]:
    name = doc.get("name", "")
    age = doc.get("age", "")
    condition = doc.get("medical_condition", "")
    print(f"  - {name}, age {age}, condition: {condition}")
```

```
Total Lipitor patients: 11140

Sample Lipitor patients (first 5):
  - aaRon MARtiNeZ, age 38, condition: Hypertension
  - rObeRt bAuer, age 68, condition: Asthma
  - ChRISToPHEr BRiGhT, age 48, condition: Asthma
  - KatHRYn StewArt, age 58, condition: Arthritis
  - dR. EilEEn thomPsoN, age 59, condition: Asthma
```

[ ]:
```python

```