TASK 6

```python
import pandas as pd

pd.set_option("display.max_columns", None)


df = pd.read_csv(
    r"C:\Users\adana\OneDrive\Pictures\Desktop\Bytewise\Task 6\dataset.csv"
)


"""We are converting every value of order_id to numeric type. error= coerce means if a value can't be convertef for
Example that value is a string then it will be converted to NaN"""

df["order_id"] = pd.to_numeric(df["order_id"], errors="coerce")

# Then in order_id column every Nan value is removed

df = df.dropna(subset=["order_id"])

# The remaining numeric values are converted into integer datatype

df["order_id"] = df["order_id"].astype(int)

"""df["product_id"] != 0 will produce a list of boolean values
and will index it to dataframe where there is true value for this condition"""
df = df[df["product_id"] != 0]

# Here the any value that is greater than 1500 will be set to 1500
df["amount"] = df["amount"].clip(upper=1500)

# This will drop null values from status
df = df.dropna(subset=["status"])
print(df)
```

ETL VS ELT

1. ETL (EXTRACT, TRANSFORM AND LOAD)

In ETL, data is extracted first and then transformed into desired formats and then loaded into the destination system.

- **USE CASE:**

  ETL is commonly used when the data needs to be cleaned and transformed before it is loaded in the destination system. It happens when the transformation process is complex and the destination system has strict schema requirements.

2. ELT (EXTRACT, LOAD AND TRANSFORM)

In ELT, data is extracted and directly loaded into the destination system and then transformation is done within the target system.

- **USE CASE:**

  It is used when dealing with large data where transformation can be applied efficiently within the destination system. It allows for flexibility and more scalable data processing.

Batch VS Streamline Pipeline

1. Batch Pipeline

In batch processing, the data is collected, processed and stored in batches. The processing happens at scheduled intervals.

- **USE CASE:**

  Batch processing is used in situations when real-time data processing and immediate insights are not required.

2. Streamline Pipeline

In streamline pipelining, data is continuously collected and processed in real-time as it flows into the system. The processing is done as soon as data arrives.

- **USE CASE:**

  It is essential when real time data insights and immediate data analysis are required.

Use-Case Scenario

Imagine a retail company wants to analyze customer behavior to optimize their marketing campaigns. They collect data from various sources like online transactions, in-store purchases, and social media interactions.

- **Solution**

  - **ELT with Streaming Process**

    The company extract data from various sources and directly load it into a cloud-based data warehouse.
    Data is processed in real-time as it is ingested. This allows the company to immediately analyze customer behavior and adjust marketing strategies.

- **Why ELT with Streaming Process is the best solution**

  1) **Real-Time Insights**

     In the retail environment, real-time insights are invaluable for quickly adjusting the strategies based on current customer behavior.

  2) **Scalability**

     The cloud-based warehouse can easily scale to handle increasing volumes of data with compromising performance.

  3) **Flexibility**

     ELT allows for more flexible transformation especially in cases when there are rapid changes in data sources