In [50]:

```python
import pandas as pd
import numpy as np
import csv
import time
import yaml
```

In [21]:

```python
start = time.time()

df_pd = pd.read_csv('Books_rating.csv')

end = time.time()
print(end - start)
```

```
28.311389207839966
```

In [23]:

```python
start = time.time()

data = []
with open("Books_rating.csv", "r") as csvfile:
    reader_variable = csv.reader(csvfile, delimiter=",")
    for row in reader_variable:
        data.append(row)

end = time.time()
print(end - start)
```

```
48.513229846954346
```

In [41]:

```
df_pd['Title'] = df_pd['Title'].apply(lambda x: ''.join(filter(str.i
df_pd.head()
```

Out[41]:

	Id	Title	Price	User_id	profileName	review/helpfulness
0	1882931173	itsonlyartifitswellhung	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7
1	0826414346	drseussamericanicon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10
2	0826414346	drseussamericanicon	NaN	A3UH4UZ4RSVO82	John Granger	10/11
3	0826414346	drseussamericanicon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7
4	0826414346	drseussamericanicon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3

In [43]:

```
df_pd.columns
```

Out[43]:

```
Index(['Id', 'Title', 'Price', 'User_id', 'profileName', 'review/helpfu
lness',
       'review/score', 'review/time', 'review/summary', 'review/text'],
      dtype='object')
```

In [44]:

```
%%writefile file.yaml
file_type: csv
dataset_name: Amazon Review
file_name: Books_rating
inbound_delimiter: ","
outbound_delimiter: "|"
skip_leading_rows: 1
columns:
  - Id
  - Title
  - Price
  - User_id
  - profileName
  - review/helpfulness
  - review/score
  - review/time
  - review/summary
  - review/text

```

```
Writing file.yaml
```

In [54]:

```
with open('file.yaml') as f:
    my_dict = yaml.safe_load(f)
my_dict['columns']
```

Out[54]:

```
['Id',
 'Title',
 'Price',
 'User_id',
 'profileName',
 'review/helpfulness',
 'review/score',
 'review/time',
 'review/summary',
 'review/text']
```

In [57]:

```
import gzip

text_data = "\n".join("|".join(row) for row in data)

with gzip.open('output_file.gz', 'wt') as f:
    f.write(text_data)
```

In [58]:

```
import os
```

In [60]:

```
size_bytes = os.path.getsize('output_file.gz')
size_bytes
```

Out[60]:

```
1054180627
```

In [62]:

```
df_pd.shape
```

Out[62]:

```
(3000000, 10)
```

Summary

Number of Columns: 10

Number of Rows: 3000000

gz File Size: 1054180627 bytes

```
In [ ]:
```