HPDP Assignment 2

MasterData

Kek Jesslyn (A22EC0057)
Tan Jun Yuan (A22EC0107)

Task1: Dataset Selection

1. The dataset we choose is Amazon Book Review

- Source: Kaggle (https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews)
- Size: 2.86 GB
- Domain: E-commerce (Online Retail)
- Number of Record: 3000000 rows x 10 columns

In [2]:
```
!pip install opendatasets
import opendatasets as od

od.download(
    "https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews")
```

```
Requirement already satisfied: opendatasets in /usr/local/lib/python3.11/dist-packages (0.1.22)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from opendatasets) (4.67.1)
Requirement already satisfied: kaggle in /usr/local/lib/python3.11/dist-packages (from opendatasets) (1.7.4.5)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from opendatasets) (8.2.1)
Requirement already satisfied: bleach in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (6.2.0)
Requirement already satisfied: certifi>=14.05.14 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2025.4.26)
Requirement already satisfied: charset-normalizer in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (3.4.2)
Requirement already satisfied: idna in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (3.10)
Requirement already satisfied: protobuf in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (5.29.4)
Requirement already satisfied: python-dateutil>=2.5.3 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.9.0.post0)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (8.0.4)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.32.3)
Requirement already satisfied: setuptools>=21.0.0 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (75.2.0)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (1.17.0)
Requirement already satisfied: text-unidecode in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (1.3)
Requirement already satisfied: urllib3>=1.15.1 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.4.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (0.5.1)
Skipping, found downloaded files in "./amazon-books-reviews" (use force=True to force download)
```

In [3]:
```
import pandas as pd
import polars as pl
import dask.dataframe as dd
import re
import time
import psutil
import tracemalloc

file =('amazon-books-reviews/Books_rating.csv')
```

In [4]:
```
def performance_measured(total_rows,execution_time,throughput,current,peak):
  print("==================== Performance ====================\n")
  print(f"Total rows processed: {total_rows}")
  print(f"Code Execution time: {execution_time:.4f} seconds")
  print(f"Throughput: {throughput:.2f} rows per second")
  print(f"Current memory usage: {current / 10**6:.4f} MB")
  print(f"Peak memory usage: {peak / 10**6:.4f} MB")

  cpu_usage = psutil.cpu_percent(interval=1)
  print(f"CPU usage: {cpu_usage}%")

  print("=====================================================")

  print("\nTotal time for this cell(Including time to display the performance):")
```

Task 2: Load and Inspect Data

In [5]:
```
%%time

tracemalloc.start()
start_time = time.perf_counter()

#Load dataset
df_whole = (pd.read_csv(file))

#Display dataset
display(df_whole.head(10))

#Display number of rows and column of dataset
print(f"Total rows: {df_whole.shape[0]}")
print(f"Total columns: {df_whole.shape[1]}\n\n\n")
```

```
current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = df_whole.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...
5	0826414346	Dr. Seuss: American Icon	NaN	A2F6NONFUDB6UK	Malvin	2/2	4.0	1127174400	One of America's greatest creative talents	"Dr. Seuss: American Icon" by Philip Nel is a ...
6	0826414346	Dr. Seuss: American Icon	NaN	A14OJS0VWMOSWO	Midwest Book Review	3/4	5.0	1100131200	A memorably excellent survey of Dr. Seuss' man...	Theodor Seuss Giesel was best known as 'Dr. Se...
7	0826414346	Dr. Seuss: American Icon	NaN	A2RSSXTDZDUSH4	J. Squire	0/0	5.0	1231200000	Academia At It's Best	When I recieved this book as a gift for Christ...
8	0826414346	Dr. Seuss: American Icon	NaN	A25MD5I2GUIW6W	J. P. HIGBED "big fellow"	0/0	5.0	1209859200	And to think that I read it on the tram!	Trams (or any public transport) are not usuall...
9	0826414346	Dr. Seuss: American Icon	NaN	A3VA4XFS5WNJO3	Donald Burnside	3/5	4.0	1076371200	Fascinating account of a genius at work	As far as I am aware, this is the first book-l...

```
Total rows: 3000000
Total columns: 10



==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 55.5354 seconds
Throughput: 54019.62 rows per second
Current memory usage: 3232.8411 MB
Peak memory usage: 3856.6924 MB
CPU usage: 4.0%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 51 s, sys: 4.38 s, total: 55.3 s
Wall time: 58.9 s
```

In [6]:

```
%%time

tracemalloc.start()
start_time = time.perf_counter()

print("==================== Inspect of Data ====================\n")

print(f"Shape: {df_whole.shape}\n")

df_whole.info()

print("\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = df_whole.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

```
==================== Inspect of Data ====================

Shape: (3000000, 10)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000000 entries, 0 to 2999999
Data columns (total 10 columns):
 #   Column              Dtype  
---  ------              -----  
 0   Id                  object 
 1   Title               object 
 2   Price               float64
 3   User_id             object 
 4   profileName         object 
 5   review/helpfulness  object 
 6   review/score        float64
 7   review/time         int64  
 8   review/summary      object 
 9   review/text         object 
dtypes: float64(2), int64(1), object(7)
memory usage: 228.9+ MB



==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 0.0254 seconds
Throughput: 118303611.39 rows per second
Current memory usage: 0.0737 MB
Peak memory usage: 0.0856 MB
CPU usage: 38.2%
=======================================================

Total time for this cell(Including time to display the performance):
CPU times: user 27.7 ms, sys: 2.1 ms, total: 29.8 ms
Wall time: 1.03 s
```

Task 3: Apply Big Data Handling Strategies

Strategy 1: Load Less Data

In [7]:
```
%%time

tracemalloc.start()
start_time = time.perf_counter()

# Load only required column
df_less = pd.read_csv(file, usecols = ["Id", "Title", "Price", "User_id",
                                       "review/helpfulness", "review/score",
                                         "review/time"])

display(df_less.head(10))
print(f"Total rows: {df_less.shape[0]}")
print(f"Total columns: {df_less.shape[1]}\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = df_less.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

	Id	Title	Price	User_id	review/helpfulness	review/score	review/time
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	7/7	4.0	940636800
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	10/10	5.0	1095724800
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	10/11	5.0	1078790400
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	7/7	4.0	1090713600
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	3/3	4.0	1107993600
5	0826414346	Dr. Seuss: American Icon	NaN	A2F6NONFUDB6UK	2/2	4.0	1127174400
6	0826414346	Dr. Seuss: American Icon	NaN	A14OJS0VWMOSWO	3/4	5.0	1100131200
7	0826414346	Dr. Seuss: American Icon	NaN	A2RSSXTDZDUSH4	0/0	5.0	1231200000
8	0826414346	Dr. Seuss: American Icon	NaN	A25MD5I2GUIW6W	0/0	5.0	1209859200
9	0826414346	Dr. Seuss: American Icon	NaN	A3VA4XFS5WNJO3	3/5	4.0	1076371200

```
Total rows: 3000000
Total columns: 7



==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 23.3248 seconds
Throughput: 128618.50 rows per second
Current memory usage: 334.9451 MB
Peak memory usage: 742.9210 MB
CPU usage: 13.6%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 23.2 s, sys: 1.06 s, total: 24.2 s
Wall time: 25.3 s
```

2. Strategy 2: Chunking

In [9]:
```
%%time

tracemalloc.start()
start_time = time.perf_counter()

#Perform chunking
df_chunk = pd.read_csv(file, chunksize=500_000)
total_rows = 0

for chunk in df_chunk:
  total_rows += chunk.shape[0]
  display(chunk.head(10))
  print(f"Total rows: {chunk.shape[0]}")
  print(f"Total columns: {chunk.shape[1]}\n")

print("\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...
5	0826414346	Dr. Seuss: American Icon	NaN	A2F6NONFUDB6UK	Malvin	2/2	4.0	1127174400	One of America's greatest creative talents	"Dr. Seuss: American Icon" by Philip Nel is a ...
6	0826414346	Dr. Seuss: American Icon	NaN	A14OJS0VWMOSWO	Midwest Book Review	3/4	5.0	1100131200	A memorably excellent survey of Dr. Seuss' man...	Theodor Seuss Giesel was best known as 'Dr. Se...
7	0826414346	Dr. Seuss: American Icon	NaN	A2RSSXTDZDUSH4	J. Squire	0/0	5.0	1231200000	Academia At It's Best	When I recieved this book as a gift for Christ...
8	0826414346	Dr. Seuss: American Icon	NaN	A25MD5I2GUIW6W	J. P. HIGBED "big fellow"	0/0	5.0	1209859200	And to think that I read it on the tram!	Trams (or any public transport) are not usuall...
9	0826414346	Dr. Seuss: American Icon	NaN	A3VA4XFS5WNJO3	Donald Burnside	3/5	4.0	1076371200	Fascinating account of a genius at work	As far as I am aware, this is the first book-l...

```
Total rows: 500000
Total columns: 10
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
500000	B000MOOAJG	Atlas Shrugged	NaN	NaN	NaN	13/26	1.0	928713600	not the best study of independence	People who read and adore this books would be ...
500001	B000MOOAJG	Atlas Shrugged	NaN	A1ALOIPRR8Q06J	2 cents "meaningless memes"	5/12	2.0	1207094400	ME, ME, ME	You can plod through this bloated tome or just...
500002	B000MOOAJG	Atlas Shrugged	NaN	A2VZE4TSUGF4OZ	K. Jimmerson	5/12	5.0	1197244800	STILL A GOOD READ!	A Definite read for freedom thinkers.If you ne...
500003	B000MOOAJG	Atlas Shrugged	NaN	A19IOUW07Z3JXD	Need more time	5/12	4.0	1188950400	Greed is Good - Where would Gordon Gekko be in...	50 year anniversary edition.Very long and the ...
500004	B000MOOAJG	Atlas Shrugged	NaN	A14EQM8HQPEOU	J. Tetreault "J. Tetreault"	53/97	1.0	1171411200	"Philosophy" for the Stupid	Now as a serious student of Philosophy, I look...
500005	B000MOOAJG	Atlas Shrugged	NaN	A1RL4AA7Q7409B	DoubleA99	14/28	2.0	994809600	Philosophies are Narrow, Book is Poorly Writte...	Upon reading Atlas Shrugged, I expected the bo...
500006	B000MOOAJG	Atlas Shrugged	NaN	A4GNDO30ABOHN	Jonathan Lund	15/30	2.0	1078963200	ehhh	There are a few reasons for which I will regar...
500007	B000MOOAJG	Atlas Shrugged	NaN	NaN	NaN	15/30	4.0	924566400	Pretty good, but her philosophy is kinda whack.	I'm a thirteen year-old seventh grader and I g...
500008	B000MOOAJG	Atlas Shrugged	NaN	A2DKTZMMG3JHN4	K. Burns	6/14	3.0	1329436800	Greed is Not Good	I had really hoped to enjoy this book more tha...
500009	B000MOOAJG	Atlas Shrugged	NaN	A3D6A6TLK9S60N	R. David Roe	6/14	5.0	1009152000	Who is John Galt? He wasn't in the last election.	As novels go, "Atlas Shrugged" is no...

```
Total rows: 500000
Total columns: 10
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
1000000	061813512X	The Complete Meat Cookbook	23.1	A2R8CRV7CHXONE	Michael D. Herrington	0/1	5.0	1287273600	Good book on meat	I liked the complete description of beef lamb ...
1000001	061813512X	The Complete Meat Cookbook	23.1	A1Z37LAM5BP52M	S. Smith	7/7	5.0	1102723200	Everything you'd ever need to know about meat	With all of the different cuts of meat availab...
1000002	061813512X	The Complete Meat Cookbook	23.1	A17QTHJESQZEX0	"beccamar"	5/5	5.0	1007683200	Hasn't Failed Me Yet!	...A couple of the recipes may seem outrageous...
1000003	061813512X	The Complete Meat Cookbook	23.1	NaN	NaN	5/5	5.0	1041379200	I've never eaten so much meat!	I was never a big meat eater, but my husband i...
1000004	061813512X	The Complete Meat Cookbook	23.1	NaN	NaN	4/4	5.0	995673600	DON'T MISS THIS ONE	I was shocked to see the few negative reviews....
1000005	061813512X	The Complete Meat Cookbook	23.1	A9UBVU4IPQ1TN	Marilou R. Guy "cookbookchic"	3/3	5.0	1011312000	The best meat cookbook	Everything I have made from this cookbook has ...
1000006	061813512X	The Complete Meat Cookbook	23.1	A2BR3ZWRL87J6S	D. Siewe	3/3	5.0	1214956800	Best of the Best	This is one of the finest cookbooks for anyone...
1000007	061813512X	The Complete Meat Cookbook	23.1	A2SUK9OWVNW9KU	Paula Ray	3/3	5.0	1031788800	The only guide you'll ever need.	This not only gives you recipes; it gives you ...
1000008	061813512X	The Complete Meat Cookbook	23.1	A1QET4646XRTZR	ao_art	2/2	3.0	1320624000	Great Book, But Needs More Pictures	"The Complete Meat Cookbook" has a wonderful a...
1000009	061813512X	The Complete Meat Cookbook	23.1	A2E0M05OFHGUNX	CaGirl	2/2	5.0	1289260800	Essential book for every kitchen! A must have!	I purchased this book for myself a few years a...

```
Total rows: 500000
Total columns: 10
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
1500000	1587249332	Black Wind: A Dirk Pitt Novel	NaN	AFNUCXXEIE4PA	I. K. Bradford	5/8	2.0	1158192000	What A Disapointment	I have always eagerly awaited each Clive Cussl...
1500001	1587249332	Black Wind: A Dirk Pitt Novel	NaN	AO7SQ1SQNTJGR	Eric W. Altmann "Eric"	5/8	1.0	1127520000	The most poorly written book I have ever read.	I have read many of Cussler's books and enjoye...
1500002	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A2OP1HD9RGX5OW	Jedidiah Palosaari "Not My Real Name"	3/5	1.0	1132531200	I was expecting more	I like Dirk Pitt. And it's not just because I ...
1500003	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A1YW4XYIA94KO9	Ravenous Reader "Literature Lover"	3/5	2.0	1113436800	Poorly, Poorly Written Effort	I am a fan of Clive Cussler and have been from...
1500004	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A2M22WHRGMZKFP	Eric H. Carpenter	1/1	3.0	1105920000	Too Many Dirks	This is a Dirk Pitt adventure, a Dirk Pitt Jr....
1500005	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A2EYCL2ZT0PSTX	Timothy J. Kindler	1/1	5.0	1105142400	The beginning of the passing of the torch	Clive Cussler has been pairing up with other a...
1500006	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A1EPBT1YNJQHAC	A. Peterson "AJ"	3/4	2.0	1119398400	repetitive	What happened to Cussler? Some of his old book...
1500007	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A2MDLL1V5GXLF3	Kilarney	3/4	1.0	1113868800	It's just getting too old	I have always been a big fan of Clive Cussler,...
1500008	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A130D2W1AQ65PF	Lawrence J. Kennedy	9/13	2.0	1107388800	Black Wind Review	I've read many of the Dirk Pitt novels and thi...
1500009	1587249332	Black Wind: A Dirk Pitt Novel	NaN	A25E44CFFC4B7T	John R. Linnell	8/12	2.0	1107388800	It's an ill wind that doesn't blow some good...	A word of explanation to Amazon readers about ...

```
Total rows: 500000
Total columns: 10
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
2000000	0553527487	Winter Solstice	NaN	NaN	NaN	2/5	4.0	980208000	Winter Solstice	I just finished Winter Solstice by Rosemund Pi...
2000001	0553527487	Winter Solstice	NaN	AFVQZQ8PW0L	Harriet Klausner	10/19	4.0	965260800	Entertaining holiday drama	Though only a sexagenarian, London actress Elf...
2000002	0553527487	Winter Solstice	NaN	A26HTU8E5SH5ES	Jenni Froelich	1/4	3.0	982022400	Interesting, but TOO long.	Before I finished this book I expected to give...
2000003	0553527487	Winter Solstice	NaN	AAXKX7DTFMB8X	Anonymous	1/4	3.0	972777600	A let-down for fans of The Shell Seekers	I would not recommend this book to anyone else...
2000004	0553527487	Winter Solstice	NaN	A2M09HQ94GWQOI	"govtlwyr"	9/18	1.0	975628800	Poorly written, plodding and predictable. . . .	This is the first book of Ms. Pilcher's I have...
2000005	0553527487	Winter Solstice	NaN	A3FYDR8WLXFEB7	J. Duetsch "bookworm"	4/10	1.0	996451200	winter solstice	Other reveiwers have described plot. I want to...
2000006	0553527487	Winter Solstice	NaN	ASSEMF90ZNJV3	Dr. Jane Branam "powerpathtolove"	2/7	1.0	1235088000	Don't Get It	I am amazed at how many people like this book....
2000007	0553527487	Winter Solstice	NaN	A1BNV503E6EINC	TAI CHI "picky"	0/4	2.0	1307836800	Just couldn't get into the book	Maybe it's the whole ENGLISH thing that's hard...
2000008	038512516X	New Games Book	NaN	A25S0HBNKBTM4C	Suzanne Crooker	8/8	5.0	972777600	Playing for Fun	This book is an excellent resource for anyone ...
2000009	038512516X	New Games Book	NaN	A5ZELOQA2OY7V	silky69	1/1	5.0	1353196800	Good Old Fashion FUN!! No Need for Technology....	This book is loaded with all sort of fun games...

```
Total rows: 500000
Total columns: 10
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
2500000	0440236223	When Good Kids Kill	NaN	A386ZUYUNQXHO7	FGarrett	10/13	5.0	957571200	See what the FBI Law Enforcement Bulletin says...	This is a fascinating, frightening read. The F...
2500001	0159006910	Journey Through the Old Testament	NaN	A2YSKFELXX8U0R	Carolina Melero	0/0	5.0	1265932800	AWSOME	THis book was for my son's book report. He enj...
2500002	0159006910	Journey Through the Old Testament	NaN	A1P1ATE0F9PTMI	Dr. Eric Breure	2/6	1.0	1127347200	Journey Through the Old Testament	Did not receive the book, have not had any res...
2500003	B00088ZNEM	Lighted windows	NaN	NaN	NaN	10/10	4.0	989971200	'Lighted Windows' lightens your heart!	A good old fashioned romance. Runaway bride di...
2500004	B00088ZNEM	Lighted windows	NaN	NaN	NaN	4/4	4.0	989971200	'Lighted Windows' lightens your heart!	A good old fashioned romance. Runaway bride di...
2500005	038525556X	Like Color to the Blind	NaN	NaN	NaN	9/9	5.0	907718400	Finding and laboriously sticking to the true s...	"Like Color to the Blind" is the thi...
2500006	038525556X	Like Color to the Blind	NaN	A2BC66A3U8S90Y	Benjamin Drasin	7/7	5.0	955238400	A powerful tale of love and humanity	Before I read this book I didn't know anything...
2500007	038525556X	Like Color to the Blind	NaN	NaN	NaN	2/2	4.0	969148800	Donna Williams is the best	again, Donna Williams takes us through the eye...
2500008	038525556X	Like Color to the Blind	NaN	NaN	NaN	2/2	5.0	857779200	A must read for anyone new to the world of rel...	Having a relationship of any kind can be a jou...
2500009	038525556X	Like Color to the Blind	NaN	A3PRGCC4N6DF1E	Nancy A.	0/0	5.0	1338854400	Third one of Donna Williams books I read	I read Nobody Nowhere, Somebody Somewhere, and...

```
Total rows: 500000
Total columns: 10



==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 58.7441 seconds
Throughput: 51068.93 rows per second
Current memory usage: 1075.6776 MB
Peak memory usage: 1190.1397 MB
CPU usage: 4.4%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 58.8 s, sys: 1.44 s, total: 1min
Wall time: 1min 1s
```

Strategy 3: Optimize Data Types

In [10]:
```
%%time

tracemalloc.start()
start_time = time.perf_counter()

def optimizedDType(df):

    df_optimized = df.copy()

    for col in df_optimized.columns:
        col_dtype = df_optimized[col].dtype

        if pd.api.types.is_integer_dtype(col_dtype):
            df_optimized[col] = pd.to_numeric(df_optimized[col], downcast='integer')

        elif pd.api.types.is_float_dtype(col_dtype):
            df_optimized[col] = pd.to_numeric(df_optimized[col], downcast='float')

    return df_optimized

df_optimized = pd.read_csv(file)

print("Before Optimization:")
df_optimized.info()
print("\n")
df_optimized = optimizedDType(df_optimized)

print("After Optimization:")
df_optimized.info()
print("\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = df_optimized.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

```
Before Optimization:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000000 entries, 0 to 2999999
Data columns (total 10 columns):
 #   Column              Dtype
---  ------              -----
 0   Id                  object
 1   Title               object
 2   Price               float64
 3   User_id             object
 4   profileName         object
 5   review/helpfulness  object
 6   review/score        float64
 7   review/time         int64
 8   review/summary      object
 9   review/text         object
dtypes: float64(2), int64(1), object(7)
memory usage: 228.9+ MB


After Optimization:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000000 entries, 0 to 2999999
Data columns (total 10 columns):
 #   Column              Dtype
---  ------              -----
 0   Id                  object
 1   Title               object
 2   Price               float32
 3   User_id             object
 4   profileName         object
 5   review/helpfulness  object
 6   review/score        float32
 7   review/time         int32
 8   review/summary      object
 9   review/text         object
dtypes: float32(2), int32(1), object(7)
memory usage: 194.5+ MB



==================== Performance ====================
Total rows processed: 3000000
Code Execution time: 58.0145 seconds
Throughput: 51711.24 rows per second
Current memory usage: 3436.6469 MB
Peak memory usage: 3856.6556 MB
CPU usage: 30.6%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 57.3 s, sys: 4.35 s, total: 1min 1s
Wall time: 1min 3s
```

Strategy 4: Apply random sampling to reduce the dataset size for fast prototyping

In [5]:

```
%%time

tracemalloc.start()
start_time = time.perf_counter()

#Random sampling 10 percent of the dataset
df_sampled = pd.read_csv(file).sample(frac=0.1, random_state=42)

display(df_sampled.head(10))
print(f"Total rows: {df_sampled.shape[0]}")
print(f"Total columns: {df_sampled.shape[1]}\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = df_sampled.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
2945667	B0006CR6U4	A dictionary of the Targumim, the Talmud Babli...	NaN	A303XPDO694V6X	Ariel	2/6	4.0	1122163200	Jastrow	Jastrow made a great workthis dictionary can h...
2352586	0897166159	Espresso Coffee: Professional Techniques	NaN	A3780H4TM9RMB8	David barnes	0/1	2.0	1356912000	NOT the book	Extremely disappointed by the SHORT length and...
1531260	0736693408	The First King of Shannara (The Sword of Shann...	NaN	A1AX6VPDQQZDPV	M Carlton	4/4	5.0	1105574400	Great (what do you expect?)	This, like all of Brook's Shannara series book...
941910	0395051029	Wuthering Heights (Riverside editions)	NaN	A35RQKCCCQ62O0	LadyJ	0/0	4.0	1353888000	Satisfied	I enjoyed this classic. I didn't know the stor...
2582125	4770016050	A Cat, a Man, and Two Women (Japans Modern Wri...	NaN	A2IJQDE1I4SIJT	David C. Arnold "master D"	1/2	5.0	1167955200	Ordered 09/02/2006, still on backorder	I would love to read this book. Have accepted ...
790281	B000GP9E9C	More Than Human	NaN	A3P92F9JAZQPIE	F Funney "CatMan"	1/2	5.0	1265846400	gestalt!	bless you! lol... sorry, but you just have to ...
1452582	B000N2HCQU	The Checklist: How to Identify True Medical Ad...	NaN	A178LDK1GMX9M6	A. Cacioli	5/6	5.0	1169683200	Well Done!!!!	Thank you, Dr. Manny. This book was such an in...
628780	B0007EW3SG	Black lamb and grey falcon: A journey through ...	NaN	NaN	NaN	8/9	5.0	992995200	Great Insight into Balkans	For anyone living in or traveling to the Balka...
2171208	0590449729	Two Crazy Pigs (Hello Reader, Level 2)	3.4	ACUX3DS8PZ1HV	Diane	3/4	5.0	1163203200	great book for kids, super photos!	I am a teacher and use this book with my Engli...
2017802	B000KAHM5Q	Love You Forever	NaN	NaN	NaN	0/1	5.0	948672000	Very sweet childrens book....a true mothers lo...	My teacher read this to our 8th grade lit. cla...

```
Total rows: 300000
Total columns: 10



==================== Performance ====================

Total rows processed: 300000
Code Execution time: 60.2045 seconds
Throughput: 4983.01 rows per second
Current memory usage: 351.7794 MB
Peak memory usage: 3856.7046 MB
CPU usage: 4.0%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 55.5 s, sys: 4.52 s, total: 1min
Wall time: 1min 1s
```

Strategy 5: Parallel Processing with Dask

In [6]:

```
%%time
tracemalloc.start()
start_time = time.perf_counter()

df_dask= dd.read_csv(file, dtype={'Id':'object'})

display(df_dask.head(10))

df_shape = df_dask.shape
n_rows = df_dask.shape[0].compute()
n_cols = df_dask.shape[1]
print(f"Total rows: {n_rows}")
print(f"Total columns: {n_cols}\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = n_rows
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka "D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...
5	0826414346	Dr. Seuss: American Icon	NaN	A2F6NONFUDB6UK	Malvin	2/2	4.0	1127174400	One of America's greatest creative talents	"Dr. Seuss: American Icon" by Philip Nel is a ...
6	0826414346	Dr. Seuss: American Icon	NaN	A14OJS0VWMOSWO	Midwest Book Review	3/4	5.0	1100131200	A memorably excellent survey of Dr. Seuss' man...	Theodor Seuss Giesel was best known as 'Dr. Se...
7	0826414346	Dr. Seuss: American Icon	NaN	A2RSSXTDZDUSH4	J. Squire	0/0	5.0	1231200000	Academia At It's Best	When I recieved this book as a gift for Christ...
8	0826414346	Dr. Seuss: American Icon	NaN	A25MD5I2GUIW6W	J. P. HIGBED "big fellow"	0/0	5.0	1209859200	And to think that I read it on the tram!	Trams (or any public transport) are not usuall...
9	0826414346	Dr. Seuss: American Icon	NaN	A3VA4XFS5WNJO3	Donald Burnside	3/5	4.0	1076371200	Fascinating account of a genius at work	As far as I am aware, this is the first book-l...

```
Total rows: 3000000
Total columns: 10


==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 77.7479 seconds
Throughput: 38586.26 rows per second
Current memory usage: 15.5786 MB
Peak memory usage: 447.7110 MB
CPU usage: 4.0%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 1min 20s, sys: 6.82 s, total: 1min 27s
Wall time: 1min 18s
```

END PART 1

Part 2: Comparing performance between different libraries

Task 1: Using Pandas library

In [8]:
```
%%time

import pandas as pd

tracemalloc.start()
start_time = time.perf_counter()

file =('amazon-books-reviews/Books_rating.csv')
pandas_df = pd.read_csv(file)

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = pandas_df.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

```
==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 55.8117 seconds
Throughput: 53752.17 rows per second
Current memory usage: 3232.6558 MB
Peak memory usage: 3856.6912 MB
CPU usage: 20.0%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 56.6 s, sys: 3.8 s, total: 1min
Wall time: 1min 2s
```

Task 2: Using Polars library

In [9]:
```
%%time

import polars as pl

tracemalloc.start()
start_time = time.perf_counter()

file =('amazon-books-reviews/Books_rating.csv')
polars_df = pl.read_csv(file)

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = polars_df.shape[0]
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

```
==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 13.4898 seconds
Throughput: 222389.96 rows per second
Current memory usage: 0.0412 MB
Peak memory usage: 0.0861 MB
CPU usage: 40.5%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 7.77 s, sys: 2.84 s, total: 10.6 s
Wall time: 14.6 s
```

Task 3: Using Dask library

In [10]:
```
%%time

import dask.dataframe as dd

tracemalloc.start()
start_time = time.perf_counter()

file =('amazon-books-reviews/Books_rating.csv')
dask_df = dd.read_csv(file, dtype={'Id':'object'})

n_rows = dask_df.shape[0].compute()
current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time

total_rows = n_rows
throughput = total_rows / execution_time

performance_measured(total_rows,execution_time,throughput,current,peak)
```

```
==================== Performance ====================

Total rows processed: 3000000
Code Execution time: 81.2019 seconds
Throughput: 36944.93 rows per second
Current memory usage: 0.1794 MB
Peak memory usage: 428.5369 MB
CPU usage: 28.0%
=====================================================

Total time for this cell(Including time to display the performance):
CPU times: user 1min 14s, sys: 8.61 s, total: 1min 23s
Wall time: 1min 22s
```

END PART 2