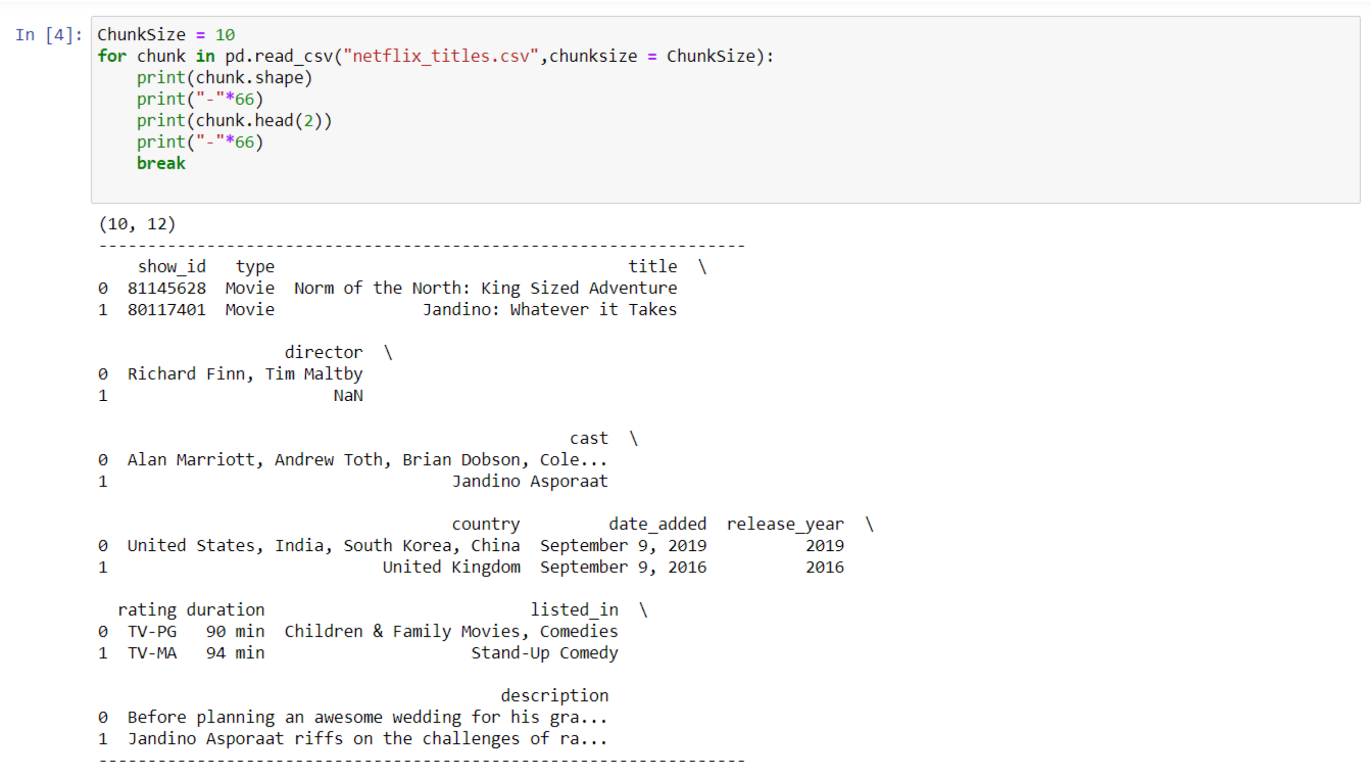
Project Name: Reading Large Files Using Chunk size

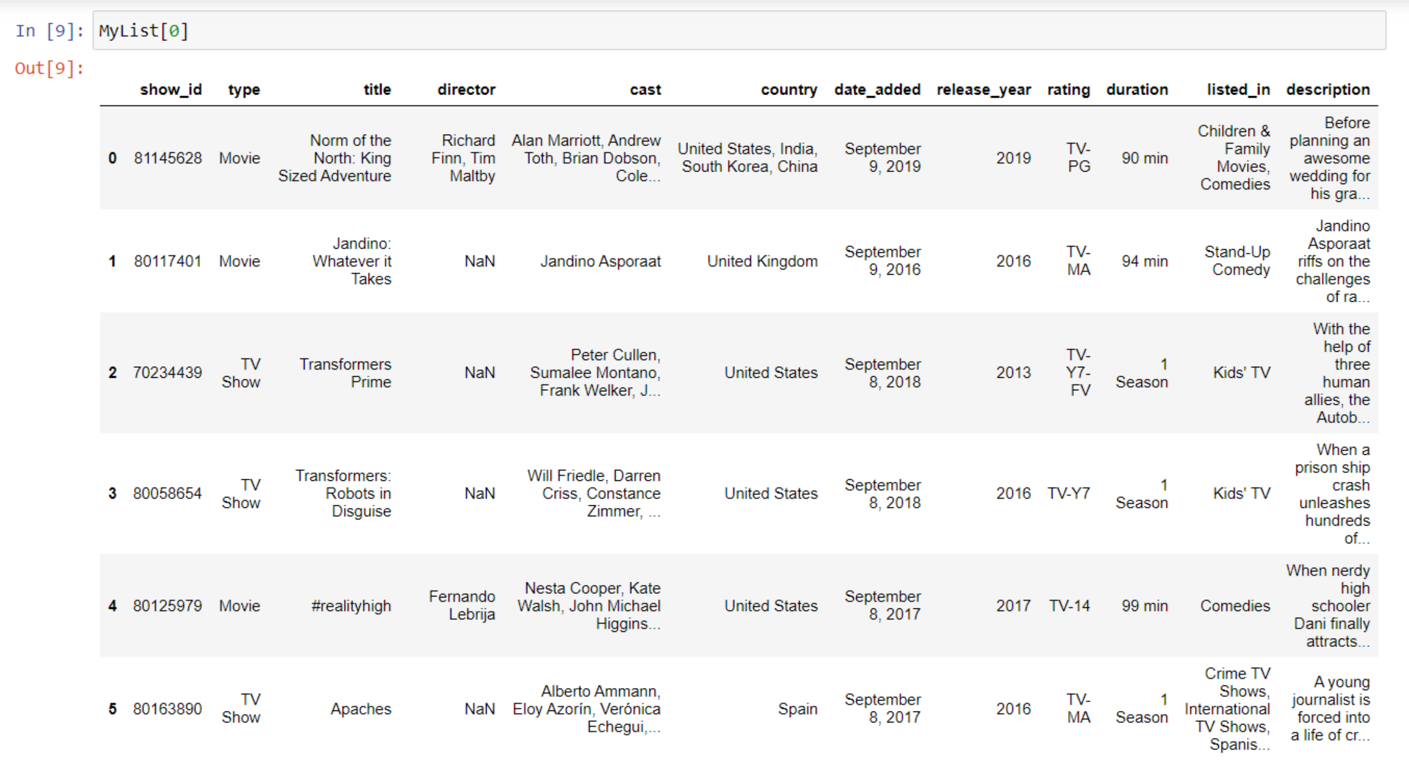
Table of Content

Demo



```
In [4]: ChunkSize = 10
        for chunk in pd.read_csv("netflix_titles.csv",chunksize = ChunkSize):
            print(chunk.shape)
            print("-"*66)
            print(chunk.head(2))
            print("-"*66)
            break
```

```
(10, 12)
------------------------------------------------------------------
    show_id   type                                    title  \
0  81145628  Movie  Norm of the North: King Sized Adventure   
1  80117401  Movie               Jandino: Whatever it Takes   

                   director  \
0  Richard Finn, Tim Maltby   
1                       NaN   

                                               cast  \
0  Alan Marriott, Andrew Toth, Brian Dobson, Cole...   
1                                  Jandino Asporaat   

                                    country         date_added  release_year  \
0  United States, India, South Korea, China  September 9, 2019          2019   
1                            United Kingdom  September 9, 2016          2016   

  rating duration                           listed_in  \
0  TV-PG   90 min  Children & Family Movies, Comedies   
1  TV-MA   94 min                     Stand-Up Comedy   

                                         description  
0  Before planning an awesome wedding for his gra...  
1  Jandino Asporaat riffs on the challenges of ra...  
------------------------------------------------------------------
```

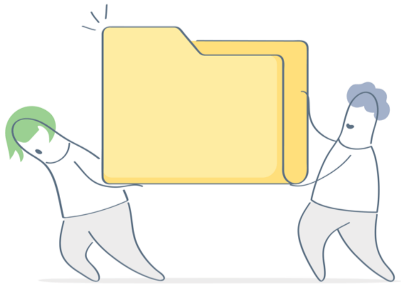


```
In [9]: MyList[0]
```

Out[9]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole...	United States, India, South Korea, China	September 9, 2019	2019	TV-PG	90 min	Children & Family Movies, Comedies	Before planning an awesome wedding for his gra...
1	80117401	Movie	Jandino: Whatever it Takes	NaN	Jandino Asporaat	United Kingdom	September 9, 2016	2016	TV-MA	94 min	Stand-Up Comedy	Jandino Asporaat riffs on the challenges of ra...
2	70234439	TV Show	Transformers Prime	NaN	Peter Cullen, Sumalee Montano, Frank Welker, J...	United States	September 8, 2018	2013	TV-Y7-FV	1 Season	Kids' TV	With the help of three human allies, the Autob...
3	80058654	TV Show	Transformers: Robots in Disguise	NaN	Will Friedle, Darren Criss, Constance Zimmer, ...	United States	September 8, 2018	2016	TV-Y7	1 Season	Kids' TV	When a prison ship crash unleashes hundreds of...
4	80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins...	United States	September 8, 2017	2017	TV-14	99 min	Comedies	When nerdy high schooler Dani finally attracts...
5	80163890	TV Show	Apaches	NaN	Alberto Ammann, Eloy Azorín, Verónica Echegui,...	Spain	September 8, 2017	2016	TV-MA	1 Season	Crime TV Shows, International TV Shows, Spanis...	A young journalist is forced into a life of cr...

Overview



This is about reading files which are large in size which may sometimes create memory errors.

There is a stark difference between large and big data.

This repository contains the code for reading large files through splitting it up into smaller chunks.

It used Pandas, os and sys libraries.

These libraries help to perform individually one particular functionality.

Data is unavoidably messy in real world.

And Pandas, is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data. Pandas objects rely heavily on Numpy objects.

Using os, means miscellaneous operating system interfaces.

sys means system-specific parameters and functions.

Data Science professionals often encounter very large data sets with hundreds of dimensions and millions of observations. So, it is one of the important skills that I am learning here. There are multiple ways to handle large data sets. It supplies precisely what we need.

Parameter essentially means the number of rows to be read into a data frame at any single time in order to fit into the local memory. Here, we are loading only some of the lines into memory at any given time. By doing this, basically we have reduced memory usage and still receive same results. The screenshot will help you to understand flow of output.

Motivation

The reason behind making is, I was baffled when I encountered an error and I couldn't read the data from csv file as my local machine has 8GB of RAM. Therefore, thought to create this one. The purpose of creating this repository is I wanted to dig deeper into Pandas, thatswhen I realized that pandas.read_csv has a parameter called chunksize. When we use argument to pandas, we get back an iterator over DataFrames rather than one single DataFrame. Though here, I have not even used that big dataset file. As I was more interested in concept rather than anything else. By building such mini project helped me to gain knowledge about other functionalities of Pandas library, which is most popular, common and even I have used almost everytime. Thatswhy Pandas is powerful.

Technical Aspect

Pandas is very efficient with small data (usually from 100MB up to 1GB) and performance is rarely a concern. Pandas has its own limitation when it comes to big data due to its algorithm and local memory constraints. Pandas module mainly works with the tabular data. It contains Data Frame and Series. Pandas is 18 to 20 times slower than Numpy. Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data.
os provides a portable way of using operating system dependent functionality. It makes it possible to automatically perform many operating system tasks. It provides functions for creating and removing a folder, fetching its contents, changing and identifying the current folder etc.
sys module provides access to some variables used by interpreter and to functions that interact strongly with the interpreter. Import sys loads the module named sys into the current namespace so that you can access the functions and anything else defined within the module using the module name. On the most common items is the list of arguments created when the program was called. sys module provides information about constants, functions and methods of the python interpreter.

Installation

Using intel core i5 9th generation with NVIDIA GFORECE GTX1650.
Windows 10 Environment Used.
Already Installed Anaconda Navigator for Python 3.x
The Code is written in Python 3.8.
If you don't have Python installed then please install Anaconda Navigator from its official site.
If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip, *python* -m pip install --*upgrade pip and press Enter.*

Run/How to Use/Steps

Keep your internet connection on while running or accessing files and throughout too.
Follow this when you want to perform from scratch.

Open Anaconda Prompt, Perform the following steps:

cd <PATH>
pip install pandas

You can also create requirement.txt file as, pip freeze > requirements.txt

run files.

If you want you can directly install packages into Jupyter Notebook Cell with ! in front.

Follow this when you want to just perform on local machine.
Download ZIP File.
Right-Click on ZIP file in download section and select Extract file option, which will unzip file.
Move unzip folder to desired folder/location be it D drive or desktop etc.
Open Anaconda Prompt, write cd <PATH> and press Enter.
eg: cd C:\Users\Monica\Desktop\Projects\Python Projects 1\5)Reading_Large_Data\Reading_Large_Files_Using_Chunksize

In Anconda Prompt, pip install -r requirements.txt to install all packages.
Open in Jupyter Notebook, <filename>.ipynb
That is,
Open in Jupyter Notebook, Reading_Large_Files_Using_Chunksize.ipynb
It takes netflix_titles.csv file as input and split it up to into smaller chunks.
Then combine those chunks into final results.
Please be careful with spellings or numbers while typing filename and easier is just copy filename and then run it to avoid any silly errors.
Note: cd <PATH>
[Go to Folder where file is. Select the path from top and right-click and select copy option and paste it next to cd one space <path> and press enter, then you can access all files of that folder] [cd means change directory]

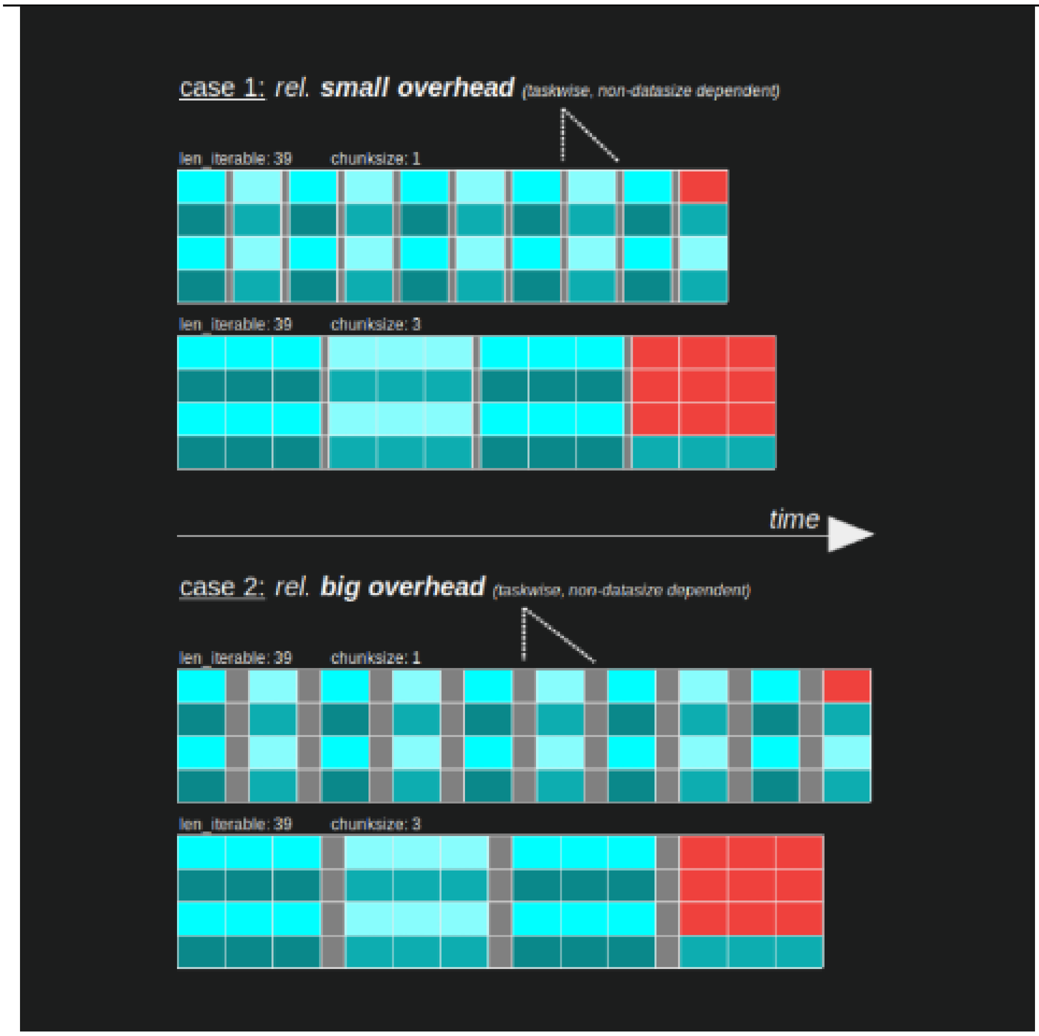
Directory Tree/Structure of Project

Folder: 5)Reading_Large_Data>Reading_Large_Files_Using_Chunksize
Reading_Large_Files_Using_Chunksize.ipynb

To Do/Future Scope

Also try with MapReduce feature.
Also add modin functionality then do chunksize.
Also try to work with AWS.

Technologies Used/System Requirement/Tech Stack


case 1: rel. small overhead (taskwise, non-datasize dependent)
len_iterable: 39
chunksize: 1
len_iterable: 39
chunksize: 3
time
case 2: rel. big overhead (taskwise, non-datasize dependent)
len_iterable: 39
chunksize: 1
len_iterable: 39
chunksize: 3

Credits

Soumilshah1995 channel