

Documentation-

The original code has only file_manager.py, I have added the following python scripts to address issues and added new operations

file_converter.py

file_copy_move_rm.py

file_parquet_reader.py

file_reader.py

file_writer.py

file_xlsx.py

gz_zip_reader.py

Testing Results:

Under Testing results, the following open issues in GitHub are resolved-

1. StopIteration Exception while reading a JSON or Parquet file

This has been handled by adding an exception, which will print 'End of the file reached' at EOF.

Here is the link to the code-

https://github.com/ankan-mazumdar/file-manager/blob/Bg-Data/file_manager/file_reader.py

```
class Reader:
    def __init__(self, file_handler):
        self.file_handler = file_handler

    def read(self):
        while True:
            line = self.file_handler.readline()
            if not line:
                print("End of file reached")
                break
            yield line
```

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_reader.py"
Enter the file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\zip_test.json
{'col1': 1, 'col2': 'A'}
{'col1': 2, 'col2': 'B'}
{'col1': 3, 'col2': 'C'}
End of file reached
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env>
```

2. Parquet read file failed due to unknown encoding

For handling the encoding-decoding failure error, I initially tried setting the encoding standard value as 'utf-8' , 'Latin-1' etc., however, none of them fixed the issue. Finally, it was a simple pandas read_parquet function that helped here. Please find the code link-

https://github.com/ankan-mazumdar/file-manager/blob/Bg-Data/file_manager/file_parquet_reader.py

Please find the test results screenshots-

```
C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data>
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_parquet_reader.py"
Enter the path of the Parquet file: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\employees.parquet

EMPLOYEE_ID  FIRST_NAME  LAST_NAME  EMAIL  PHONE_NUMBER  ...  JOB_ID  SALARY  COMMISSION_PCT  MANAGER_ID  DEPARTMENT_ID
ID
0            198      Donald    OConnell  DOCONNEL  650.507.9833  ...  SH_CLERK  2600          -         124
50
1            199      Douglas    Grant     DGRANT    650.507.9844  ...  SH_CLERK  2600          -         124
50
2            200      Jennifer  Whalen    JWHALEN    515.123.4444  ...  AD_ASST  4400          -         101
10
3            201      Michael   Hartstein  MHARTSTE  515.123.5555  ...  MK_MAN   13000         -         100
20

PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_parquet_reader.py"
Enter the path of the Parquet file: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\sample_data.parquet

name  age
0    Alice  30
1     Bob   25
2   Charlie  22

PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env>

PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_parquet_reader.py"
Enter the path of the Parquet file: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\data2.parquet

url  url_length  hostname_length  count-  count@  ...  count-www  count-digits  count-letters  count_dir  lab
0            https://www.google.com          22          14          0          0  ...          1          0          17          0
0
1            https://www.webfx.com/digital-marketing/          37          20          0          0  ...          0          0          30          3
0
2            https://www.facebook.com          24          16          0          0  ...          1          0          19          0
0
3            https://www.baidu.com          21          13          0          0  ...          1          0          16          0
0
4            https://www.wikipedia.org          25          17          0          0  ...          1          0          20          0
0
5 http://faboleena.com/js/infortis/jquery/plugin...          159          13          0          0  ...          0          21          118          12
1
6 http://faboleena.com/js/infortis/jquery/plugin...          147          13          0          0  ...          0          20          109          12
1
7            https://www.google.co.in          24          16          0          0  ...          1          0          18          0
0
8            https://www.qq.com          18          10          0          0  ...          1          0          13          0
0
9            https://www.amazon.com          22          14          0          0  ...          1          0          17          0
0

[10 rows x 11 columns]
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env>
```

3. The gzip file read operation Error: Raised OSError on reading a gzip file due to an invalid argument (Invalid characters in the file path).

The error occurs as the encoding can't recognize the characters, on researching, I found that it is Latin-1 characters rather than utf-8, hence I kept encoding ='Latin-1' in the solution code that worked perfectly. Here's the link to the code-

https://github.com/ankan-mazumdar/file-manager/blob/Bg-Data/file_manager/gz_zip_reader.py

Below is the testing result-

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\gz_zip_reader.py"
Choose an action (read, write, convert): read
Enter the file name (include the extension): C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\zip_test.csv.gz
col1,col2
1,A
2,B
3,C
```

4. Next, as proposed I have added the functionality for read-write operations for xlsx and zip files-

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_xlsx.py"
Enter the file name (include the extension): C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\data2.xlsx
.xlsx
Choose an action (read, write, convert): read
Content of XLSX file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\data2.xlsx':
      url url_length hostname_length ... count-letters count_dir label
0      https://www.google.com          22          14 ...          17          0  0
1  https://www.webfx.com/digital-marketing/ 37          20 ...          30          3  0
2      https://www.facebook.com          24          16 ...          19          0  0
3      https://www.baidu.com           21          13 ...          16          0  0
4      https://www.wikipedia.org         25          17 ...          20          0  0
5 http://faboleena.com/js/infortis/jquery/plugin... 159          13 ...         118         12  1
6 http://faboleena.com/js/infortis/jquery/plugin... 147          13 ...         109         12  1
7      https://www.google.co.in          24          16 ...          18          0  0
8      https://www.qq.com              18          10 ...          13          0  0
9      https://www.amazon.com           22          14 ...          17          0  0

[10 rows x 11 columns]
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\env>
```

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_writer.py"
Enter the file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\xlsx_write.xlsx
This is a Big Data Project content.
The file has been zipped and saved as C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\xlsx_write.xlsx.zip
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data>
```

The screenshot shows a Windows File Explorer window with the address bar set to 'This PC > Downloads > Project_BigDataCSP_554 > file-manager-master > data'. The file list contains two items: 'xlsx_write' (Microsoft Excel Worksheet, 1 KB) and 'xlsx_write.xlsx' (WinRAR ZIP archive, 1 KB). Below the file explorer, a Notepad++ window is open, displaying the content of 'xlsx_write.xlsx'. The text in the Notepad++ window is 'This is a Big Data Project content.'

5. Converter- This script can do all-round file format conversion from csv to xml, json and parquet. Parquet to csv json xml and further in this fashion. Code link-

https://github.com/ankan-mazumdar/file-manager/blob/Bg-Data/file_manager/file_converter.py

Testing results-

Csv to Parquet conversion-

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_converter.py"
Enter the source file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\employees.csv
CSV file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\employees.csv' converted to Parquet yeess.parquet
file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\employees.parquet' successfully.

BigDataCSP_554\file-manager-master\data\employees.parquet C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data>
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_parquet_reader.py"
Enter the path of the Parquet file: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\employees.parquet

EMPLOYEE_ID FIRST_NAME LAST_NAME EMAIL PHONE_NUMBER ... JOB_ID SALARY COMMISSION_PCT MANAGER_ID DEPARTMENT_ID
0 198 Donald OConnell DOCONNEL 650.507.9833 ... SH_CLERK 2600 - 124
50
1 199 Douglas Grant DGRANT 650.507.9844 ... SH_CLERK 2600 - 124
50
2 200 Jennifer Whalen JWHALEN 515.123.4444 ... AD_ASST 4400 - 101
10
3 201 Michael Hartstein MHARTSTE 515.123.5555 ... MK_MAN 13000 - 100
20
```

Json to parquet -

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_converter.py"
Enter the source file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.json
Enter the destination Parquet file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.parquet
JSON file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.json' converted to Parquet file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.parquet' successfully.
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager>

PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_parquet_reader.py"
Enter the path of the Parquet file: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.parquet

registration dtm id first_name last_name email gender ... cc country birthdate salary title comment
ts
0 1454486129000 1 Amanda Jordan ajordan0@com.com Female ... 6759521864920116 Indonesia 3/8/1971 49756.53 Internal Auditor 1F+
02
1 1454519043000 2 Albert Freeman afreeman1@is.gd Male ... Canada 1/16/1968 150280.17 Accountant IV
2 1454461771000 3 Evelyn Morgan emorgan2@altervista.org Female ... 6767119071901597 Russia 2/1/1960 144972.51 Structural Engineer
3 1454459781000 4 Denise Riley driley3@gmpg.org Female ... 3576031598965625 China 4/8/1997 90263.05 Senior Cost Accountant
4 1454475931000 5 Carlos Burns cburns4@mitbeian.gov.cn ... 5602256255204850 South Africa NaN
..
..
995 14544945459000 996 Dennis Harris dharrisrr@eepurl.com Male ... 374288806662929 Greece 7/8/1965 263399.54 Editor
996 1454519813000 997 Gloria Hamilton ghamiltonrc@rambler.ru Female ... China 4/22/1975 83183.54 VP Product Management
997 1454475740000 998 Nancy Morris nmorrisrp@ask.com ... 3553564071014997 Sweden 5/1/1979 NaN Junior Executive
998 1454467292000 999 Annie Daniels adanielsrq@squidoo.com Female ... 30424803513734 China 10/9/1991 18433.85 Editor
999 1454493138000 1000 Julie Meyer jmeyerrrr@flavors.me Female ... 374288099198540 China 222561.13

[1000 rows x 13 columns]
```

Parquet to Json-

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_converter.py"
Enter the source file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.parquet
Enter the destination file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.json
Choose conversion type (parquet_to_csv, parquet_to_xml, parquet_to_json): parquet_to_json
Parquet file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.parquet' converted to JSON file 'C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.json' successfully.
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_reader.py"
Enter the file path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\userdata1.json
```

```
{'registration_dttm': 1454519813800, 'id': 997, 'first_name': 'Gloria', 'last_name': 'Hamilton', 'email': 'ghamiltonro@rambler.ru', 'gender': 'Female', 'ip_address': '71.50.39.137', 'cc': '', 'country': 'China', 'birthdate': '4/22/1975', 'salary': 83183.54, 'title': 'VP Product Management', 'comments': ''}
{'registration_dttm': 1454475740000, 'id': 998, 'first_name': 'Nancy', 'last_name': 'Morris', 'email': 'nmorrisrp@ask.com', 'gender': '', 'ip_address': '6.188.121.221', 'cc': '3553564071014997', 'country': 'Sweden', 'birthdate': '5/1/1979', 'salary': None, 'title': 'Junior Executive', 'comments': ''}
{'registration_dttm': 1454467292800, 'id': 999, 'first_name': 'Annie', 'last_name': 'Daniels', 'email': 'adanielsrq@squidoo.com', 'gender': 'Female', 'ip_address': '97.221.132.35', 'cc': '30424883513734', 'country': 'China', 'birthdate': '10/9/1991', 'salary': 18433.85, 'title': 'Editor', 'comments': '\u200b'}
{'registration_dttm': 1454493138000, 'id': 1000, 'first_name': 'Julie', 'last_name': 'Meyer', 'email': 'jmeyerrrr@flavors.me', 'gender': 'Female', 'ip_address': '217.1.147.132', 'cc': '374288099198540', 'country': 'China', 'birthdate': '', 'salary': 222561.13, 'title': '', 'comments': ''}
End of file reached
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager>
```

6. Copy move, create remove operations- As name suggests this script basically can copy , move, make, and remove files from any mentioned path. Here is the code link-
https://github.com/ankan-mazumdar/file-manager/blob/Bg-Data/file_manager/file_copy_move_rm.py

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_copy_move_rm.py"
Choose an action (copy, move, remove, create, or exit): move
Enter the source file/folder path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\xlsx_write.xlsx
Enter the destination file/folder path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\xlsx_write.xlsx
File moved successfully from C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data\xlsx_write.xlsx to C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\xlsx_write.xlsx
Do you want to do anything else? (yes/no): no
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\data> cd ..
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master> dir
```

Directory: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master

Mode	LastWriteTime	Length	Name
d----	03-12-2023 21:55		data
d----	18-11-2023 06:38		env
d----	03-12-2023 13:17		file_manager
d----	18-11-2023 19:03		test2
-----	28-04-2020 06:55	2043	.gitignore
-a----	18-11-2023 07:05	22	abcd
-a----	03-12-2023 19:55	192504	data.zip
-a----	18-11-2023 07:19	0	file
-a----	18-11-2023 16:51	1527	gzip error.txt
-----	28-04-2020 06:55	390	README.md
-a----	05-11-2023 13:10	87	requirements.txt
-a----	18-11-2023 16:53	39437	Screenshot 2023-11-18 165235.png
-----	28-04-2020 06:55	677	setup.py
-a----	03-12-2023 19:01	35	xlsx_write.xlsx

```
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master> & C:/Python/python.exe "C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\file_manager\file_copy_move_rm.py"
Choose an action (copy, move, remove, create, or exit): remove
Enter the source file/folder path: C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\xlsx_write.xlsx
File C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master\xlsx_write.xlsx removed successfully
Do you want to do anything else? (yes/no): no
PS C:\Users\Ankan Mazumdar\Downloads\Project_BigDataCSP_554\file-manager-master>
```

7. Uploaded the scripts over EMR cluster and tested using python and Spark-

a. Csv file -

```
[hadoop@ip-172-31-46-19 file-manager-master]$ python file_manager/file_reader.py
Enter the file path: /home/hadoop/file-manager-master/data/data2.csv
OrderedDict([('url', 'https://www.google.com/'), ('url_length', '22'), ('hostname_length', '14'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '17'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'https://www.webfx.com/digital-marketing/'), ('url_length', '37'), ('hostname_length', '20'), ('count-', '0'), ('count@', '0'), ('count-https', '0'), ('count-www', '0'), ('count-digits', '0'), ('count-letters', '30'), ('count_dir', '3'), ('label', '0')])
OrderedDict([('url', 'https://www.facebook.com/'), ('url_length', '24'), ('hostname_length', '16'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '19'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'https://www.baidu.com/'), ('url_length', '21'), ('hostname_length', '13'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '16'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'https://www.wikipedia.org/'), ('url_length', '25'), ('hostname_length', '17'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '20'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'http://faboleena.com/js/infortis/jquery/plugins/minicolors/Ameli/PortailAS/appmanager/PortailAS/assure_somtc=true/018ad7352b902f40e0e37b0f53cf47fc/index_2.html'), ('url_length', '159'), ('hostname_length', '13'), ('count-', '0'), ('count@', '0'), ('count-https', '0'), ('count-www', '0'), ('count-digits', '21'), ('count-letters', '118'), ('count_dir', '12'), ('label', '1')])
OrderedDict([('url', 'http://faboleena.com/js/infortis/jquery/plugins/minicolors/Ameli/PortailAS/appmanager/PortailAS/assure_somtc=true/018ad7352b902f40e0e37b0f53cf47fc/'), ('url_length', '147'), ('hostname_length', '13'), ('count-', '0'), ('count@', '0'), ('count-https', '0'), ('count-www', '0'), ('count-digits', '20'), ('count-letters', '109'), ('count_dir', '12'), ('label', '1')])
OrderedDict([('url', 'https://www.google.co.in/'), ('url_length', '24'), ('hostname_length', '16'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '18'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'https://www.qq.com/'), ('url_length', '18'), ('hostname_length', '10'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '13'), ('count_dir', '0'), ('label', '0')])
OrderedDict([('url', 'https://www.amazon.com/'), ('url_length', '22'), ('hostname_length', '14'), ('count-', '0'), ('count@', '0'), ('count-https', '1'), ('count-www', '1'), ('count-digits', '0'), ('count-letters', '17'), ('count_dir', '0'), ('label', '0')])
[hadoop@ip-172-31-46-19 file-manager-master]$
```

b. Json file-

```
Enter the file path: /home/hadoop/file-manager-master/data/zip_test.json
{'col1': 1, 'col2': 'A'}
{'col1': 2, 'col2': 'B'}
{'col1': 3, 'col2': 'C'}
End of file reached
[hadoop@ip-172-31-46-19 file_manager]$
```

c. Parquet file reading-

```
[hadoop@ip-172-31-46-19 file_manager]$ python file_parquet_reader.py
Enter the path of the Parquet file: /home/hadoop/file-manager-master/data/userdata1.parquet
  registration_dttm  id first_name last_name  email  gender  cc  country  birthdate  salary
0  2016-02-03 07:55:29  1  Amanda  Jordan  ajordan0@com.com  Female  ...  6759521864920116  Indonesia  3/8/1971  49756.53
1  2016-02-03 17:04:03  2  Albert  Freeman  afreeman1@is.gd  Male  ...  ...  Canada  1/16/1968  150280.17
2  2016-02-03 01:09:31  3  Evelyn  Morgan  emorgan2@altervista.org  Female  ...  6767119071901597  Russia  2/1/1960  144972.51
3  2016-02-03 00:36:21  4  Denise  Riley  driley3@gmpg.org  Female  ...  3576031598965625  China  4/8/1997  90263.05
4  2016-02-03 05:05:31  5  Carlos  Burns  cburns4@mitbeian.gov.cn  ...  5602256255204850  South Africa  NaN
..  ..  ..  ..  ..  ..  ..  ..  ..  ..
995 2016-02-03 10:30:59  996  Dennis  Harris  dharrisrn@eepurl.com  Male  ...  374288806662929  Greece  7/8/1965  263399.54
996 2016-02-03 17:16:53  997  Gloria  Hamilton  ghamiltonro@rambler.ru  Female  ...  ...  China  4/22/1975  83183.54
997 2016-02-03 05:02:20  998  Nancy  Morris  nmorrisrp@ask.com  ...  3553564071014997  Sweden  5/1/1979  NaN
998 2016-02-03 02:41:32  999  Annie  Daniels  adanielsrq@squidoo.com  Female  ...  30424803513734  China  10/9/1991  18433.85
999 2016-02-03 09:52:18  1000  Julie  Meyer  jmeyerrrr@flavors.me  Female  ...  374288099198540  China  222561.13

[1000 rows x 13 columns]
[hadoop@ip-172-31-46-19 file_manager]$
```

d. Xlsx file-

```
[hadoop@ip-172-31-46-19 file_manager]$ python file_xlsx.py
Enter the file name (include the extension): /home/hadoop/file-manager-master/data/userdata1.xlsx
Choose an action (read, write, convert): read
Content of XLSX file '/home/hadoop/file-manager-master/data/userdata1.xlsx':
  registration_dttm  id first_name last_name  email  gender  cc  country  birthdate  salary  title  comments
0  2016-02-03 07:55:29  1  Amanda  Jordan  ajordan0@com.com  Female  ...  6.759522e+15  Indonesia  3/8/1971  49756.53  Internal Auditor  100.0
1  2016-02-03 17:04:03  2  Albert  Freeman  afreeman1@is.gd  Male  ...  NaN  Canada  1/16/1968  150280.17  Accountant IV  NaN
2  2016-02-03 01:09:31  3  Evelyn  Morgan  emorgan2@altervista.org  Female  ...  6.767119e+15  Russia  2/1/1960  144972.51  Structural Engineer  NaN
3  2016-02-03 00:36:21  4  Denise  Riley  driley3@gmpg.org  Female  ...  3.576032e+15  China  4/8/1997  90263.05  Senior Cost Accountant  NaN
4  2016-02-03 05:05:31  5  Carlos  Burns  cburns4@mitbeian.gov.cn  NaN  ...  5.602256e+15  South Africa  NaN  NaN
5  2016-02-03 07:12:34  6  Kathryn  White  kwhites@google.com  Female  ...  3.583136e+15  Indonesia  2/25/1983  69227.11  Account Executive  NaN
6  2016-02-03 08:33:08  7  Samuel  Holmes  sholmes6@foxnews.com  Male  ...  3.582641e+15  Portugal  12/18/1987  14247.62  Senior Financial Analyst  NaN
7  2016-02-03 06:47:06  8  Harry  Howell  hhowell17@eepurl.com  Male  ...  NaN  Bosnia and Herzegovina  3/1/1962  186469.43  Web Developer IV  NaN
8  2016-02-03 03:52:53  9  Josee  Foster  jfoster8@yelp.com  Male  ...  NaN  South Korea  3/27/1992  231067.84  Software Test Engineer I  100.0
9  2016-02-03 18:29:47  10  Emily  Stewart  estewart9@opensource.org  Female  ...  3.574254e+15  Nigeria  1/28/1997  27234.28  Health Coach IV  NaN

[10 rows x 13 columns]
[hadoop@ip-172-31-46-19 file_manager]$
```

e. Compressed file-


```
[10 rows x 13 columns]
[hadoop@ip-172-31-46-19 file_manager]$ python gz_zip_reader.py
Choose an action (read, write, convert): read
Enter the file name (include the extension): /home/hadoop/file-manager-master/data/employee.zip
Content of acbb2271e66c10a5b73aacf82ca82784-e38afe62e088394d61ed30884dd50a6826eee0a8/:

Content of acbb2271e66c10a5b73aacf82ca82784-e38afe62e088394d61ed30884dd50a6826eee0a8/employees.csv:
EMPLOYEE_ID,FIRST_NAME, LAST_NAME, EMAIL, PHONE_NUMBER, HIRE_DATE, JOB_ID, SALARY, COMMISSION_PCT, MANAGER_ID, DEPARTMENT_ID
198,Donald,OConnell,DOCONNEL,650.507.9833,21-JUN-07,SH_CLERK,2600, -,124,50
199,Douglas,Grant,DGRANT,650.507.9844,13-JAN-08,SH_CLERK,2600, -,124,50
200,Jennifer,Whalen,JWHALEN,515.123.4444,17-SEP-03,AD_ASST,4400, -,101,10
201,Michael,Hartstein,MHARTSTE,515.123.5555,17-FEB-04,MK_MAN,13000, -,100,20
202,Pat,Fay,PFAY,603.123.6666,17-AUG-05,MK_REP,6000, -,201,20
203,Susan,Mavris,SMAVRIS,515.123.7777,07-JUN-02,HR_REP,6500, -,101,40
204,Hermann,Baer,HBAER,515.123.8888,07-JUN-02,PR_REP,10000, -,101,70
205,Shelley,Higgins,SHIGGINS,515.123.8080,07-JUN-02,AC_MGR,12008, -,101,110
206,William,Gietz,WGIETZ,515.123.8181,07-JUN-02,AC_ACCOUNT,8300, -,205,110
100,Steven,King,SKING,515.123.4567,17-JUN-03,AD PRES,24000, -, -,90
101,Neena,Kochhar,NKOCHHAR,515.123.4568,21-SEP-05,AD_VP,17000, -,100,90
102,Lex,De Haan,LDEHAAN,515.123.4569,13-JAN-01,AD_VP,17000, -,100,90
103,Alexander,Hunold,AHUNOLD,590.423.4567,03-JAN-06,IT_PROG,9000, -,102,60
104,Bruce,Ernst,BERNST,590.423.4568,21-MAY-07,IT_PROG,6000, -,103,60
105,David,Austin,DAUSTIN,590.423.4569,25-JUN-05,IT_PROG,4800, -,103,60
106,Valli,Pataballa,VPATABAL,590.423.4560,05-FEB-06,IT_PROG,4800, -,103,60
107,Diana,Lorentz,DLORENTZ,590.423.5567,07-FEB-07,IT_PROG,4200, -,103,60
108,Nancy,Greenberg,NGREENBE,515.124.4569,17-AUG-02,FI_MGR,12008, -,101,100
109,Daniel,Faviet,DFAVIET,515.124.4169,16-AUG-02,FI_ACCOUNT,9000, -,108,100
110,John,Chen,JCHEN,515.124.4269,28-SEP-05,FI_ACCOUNT,8200, -,108,100
111,Ismael,Sciarra,ISCIARRA,515.124.4369,30-SEP-05,FI_ACCOUNT,7700, -,108,100
112,Jose Manuel,Urman,JMURMAN,515.124.4469,07-MAR-06,FI_ACCOUNT,7800, -,108,100
113,Luis,Popp,LPOPP,515.124.4567,07-DEC-07,FI_ACCOUNT,6900, -,108,100
114,Den,Raphaely,DRAPHEAL,515.127.4561,07-DEC-02,PU_MAN,11000, -,100,30
115,Alexander,Khoo,AKHOO,515.127.4562,18-MAY-03,PU_CLERK,3100, -,114,30
116,Shelli,Baida,SBAIDA,515.127.4563,24-DEC-05,PU_CLERK,2900, -,114,30
117,Sigal,Tobias,STOBIAS,515.127.4564,24-JUL-05,PU_CLERK,2800, -,114,30
118,Guy,Himuro,GHIMURO,515.127.4565,15-NOV-06,PU_CLERK,2600, -,114,30
```

8. Using Spark-submit command-

Issue -Spark-submit to run Python script gets stuck and never proceeds when there is user interaction/inputs are required at run time

```
[hadoop@ip-172-31-46-19 file_manager]$ spark-submit file_xlsx.py
Enter the file name (include the extension): /home/hadoop/file-manager-master/data/userdata1.xlsx
```

<https://stackoverflow.com/questions/40910869/python-script-hangs-on-input-method-when-running-spark>.

When we submit an application using spark-submit, we don't interact with Python code, but with Java one, which doesn't expect any input from stdin.

If we want to make it work we have to skip spark-submit and execute this as a Python script. We could alternatively achieve this by using pyspark or hardcode the filename in the script

9. Read the file using Pyspark

```
>>> from pyspark import SparkConf, SparkContext
>>> def read_parquet(file_path):
...     df = pd.read_parquet(file_path)
...     return df
...
>>> file_path = input("Enter the path of the Parquet file: ")
Enter the path of the Parquet file: /home/hadoop/file-manager-master/data/sample_data.parquet
>>> data_frame = read_parquet(file_path)
>>> print(data_frame)
   name  age
0  Alice   30
1   Bob    25
2 Charlie   22
>>> from pyspark import SparkConf, SparkContext
```

10. Writing parquet file and zipping it-

```
>>> import zipfile
>>> mode = 'w'
>>> # Ask the user to input the file path
>>> file_path = input("Enter the file path: ")
Enter the file path: /home/hadoop/file-manager-master/data/sample_test.parquet
>>> with open(file_path, "w") as file:
...     file.write("This is a Big Data Project content.")
...
35
>>> with open(file_path, "r") as file:
...     print(file.read())
...
This is a Big Data Project content.
>>> zip_file_path = file_path + ".zip"
>>> zipfile.ZipFile(zip_file_path, mode='w').write(file_path)
>>> print(f'The file has been zipped and saved as {zip_file_path}')
The file has been zipped and saved as /home/hadoop/file-manager-master/data/sample_test.parquet.zip
>>>
```

The file is created -

```
[hadoop@ip-172-31-46-19 data]$ ls -l /home/hadoop/file-manager-master/data/sample_test.parquet.zip
-rw-rw-r-- 1 hadoop hadoop 245 Dec  4 03:06 /home/hadoop/file-manager-master/data/sample_test.parquet.zip
```

11.Xlsx file in spark

```
[hadoop@ip-172-31-46-19 file_manager]$ spark-submit file_xlsx.py
Content of XLSX file '/home/hadoop/file-manager-master/data/userdata1.xlsx':
  registration_dttm  id  ...  title  comments
0 2016-02-03 07:55:29  1  ...  Internal Auditor  100.0
1 2016-02-03 17:04:03  2  ...  Accountant IV  NaN
2 2016-02-03 01:09:31  3  ...  Structural Engineer  NaN
3 2016-02-03 00:36:21  4  ...  Senior Cost Accountant  NaN
4 2016-02-03 05:05:31  5  ...  NaN  NaN
5 2016-02-03 07:22:34  6  ...  Account Executive  NaN
6 2016-02-03 08:33:08  7  ...  Senior Financial Analyst  NaN
7 2016-02-03 06:47:06  8  ...  Web Developer IV  NaN
8 2016-02-03 03:52:53  9  ...  Software Test Engineer I  100.0
9 2016-02-03 18:29:47  10 ...  Health Coach IV  NaN

[10 rows x 13 columns]
23/12/04 03:31:13 INFO ShutdownHookManager: Shutdown hook called
23/12/04 03:31:13 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-aadb23ce-44af-42c9-bc08-12bfd518cfd1
```