

【python】資料讀寫速度比較

大家好，

有時我們在處理資料時，

會需要先暫存下來，

後續再以python做其他運用。

(目前想到的情境是模型組，可能會需要先存下各面向因子資料或其他)

最近剛好看到一篇比較各資料儲存類型的文章，

因為儲存速度跟存下來的大小差蠻多的，

因此分享一下~~~

共比較四種資料儲存格式，

分別為 CSV、pickle、parquet、feather

其中parquet 需要先另外import 套件，

參考程式碼如下

名稱	修改日期	類型	大小
測試資料導入.ipynb	2023/2/6 下午 07:00	IPYNB 檔案	4 KB

```
In [1]: # pip install -i https://[redacted] --trusted-host [redacted] pyarrow
```

結論如下:

1. 存取速度: feather > pickle > parquet >>>> CSV
2. 讀取速度: pickle > feather > parquet >>>> CSV
3. 誰最胖!! : CSV >>>> pickle > feather > parquet

簡單來說，parquet最瘦，如果需要儲存的資料量真的太大了，可以考慮使用parquet，但如果是存取速度的話，就可以考慮feather 或 pickle，兩者差距沒有到很明顯。

```
: print('Reading and writing CSV')
%time df.to_csv('test.csv')
%time df_csv = pd.read_csv('test.csv')
```

Reading and writing CSV
Wall time: 2min 40s
Wall time: 12 s

```
: print('Reading and writing Pickle')
%time df.to_pickle('test.pickle')
%time df_pickle = pd.read_pickle('test.pickle')
```

Reading and writing Pickle
Wall time: 639 ms
Wall time: 92 ms

```
: print('Reading and writing Parquet')
%time df.to_parquet('test.parquet')
%time df_parquet = pd.read_parquet('test.parquet')
```

Reading and writing Parquet
Wall time: 1.58 s
Wall time: 281 ms

```
: print('Reading and writing Feather')
%time df.to_feather('test.feather')
%time df_feather = pd.read_feather('test.feather')
```

Reading and writing Feather
Wall time: 487 ms
Wall time: 230 ms

test.csv - 內容	test.parquet - 內容	test.pickle - 內容	test.feather - 內容
<p>一般 安全性 詳細資料 以前的版本</p> <p>test.csv</p> <p>檔案類型: CSV 檔案 (.csv) 開啟檔案: 挑選應用程式</p> <p>位置: C:\Users\esb21774\Desktop 大小: 237 MB (249,438,108 位元組) 磁碟大小: 237 MB (249,438,208 位元組)</p> <p>建立日期: 2023年2月6日, 下午 06:51:01 修改日期: 2023年2月6日, 下午 06:53:42 存取日期: 2023年2月6日, 下午 06:53:42</p> <p>屬性: <input type="checkbox"/> 唯讀(R) <input type="checkbox"/> 隱藏(H)</p> <p>確定 取消</p>	<p>一般 安全性 詳細資料 以前的版本</p> <p>test.parquet</p> <p>檔案類型: PARQUET 檔案 (.parquet) 開啟檔案: 挑選應用程式</p> <p>位置: C:\Users\esb21774\Desktop 大小: 32.8 MB (34,427,866 位元組) 磁碟大小: 32.8 MB (34,430,976 位元組)</p> <p>建立日期: 2023年2月6日, 下午 06:48:28 修改日期: 2023年2月6日, 下午 07:00:03 存取日期: 2023年2月6日, 下午 07:00:03</p> <p>屬性: <input type="checkbox"/> 唯讀(R) <input type="checkbox"/> 隱藏(H)</p> <p>確定 取消</p>	<p>一般 安全性 詳細資料 以前的版本</p> <p>test.pickle</p> <p>檔案類型: PICKLE 檔案 (.pickle) 開啟檔案: 挑選應用程式</p> <p>位置: C:\Users\esb21774\Desktop 大小: 81.0 MB (85,001,824 位元組) 磁碟大小: 81.0 MB (85,004,288 位元組)</p> <p>建立日期: 2023年2月6日, 下午 06:54:07 修改日期: 2023年2月6日, 下午 06:54:08 存取日期: 2023年2月6日, 下午 06:54:08</p> <p>屬性: <input type="checkbox"/> 唯讀(R) <input type="checkbox"/> 隱藏(H)</p> <p>確定 取消</p>	<p>一般 安全性 詳細資料 以前的版本</p> <p>test.feather</p> <p>檔案類型: FEATHER 檔案 (.feather) 開啟檔案: 挑選應用程式</p> <p>位置: C:\Users\esb21774\Desktop 大小: 48.8 MB (51,268,170 位元組) 磁碟大小: 48.8 MB (51,269,632 位元組)</p> <p>建立日期: 2023年2月6日, 下午 06:54:36 修改日期: 2023年2月6日, 下午 07:00:04 存取日期: 2023年2月6日, 下午 07:00:04</p> <p>屬性: <input type="checkbox"/> 唯讀(R) <input type="checkbox"/> 隱藏(H)</p> <p>確定 取消</p>