music

February 11, 2025

[3]:
```
# Import
import pandas as pd
```

[4]:
```
# Read Files

# Path to your Parquet file
parquet_file = '/Users/henryding/Documents/Seminar/Data/output_data.parquet'

# Read the Parquet file into a DataFrame
df = pd.read_parquet(parquet_file, engine='pyarrow')  # You can also use␣
 ↪'fastparquet'

# Display the first few rows of the DataFrame
display(df.head(1000))
num_rows = df.shape[0]
print(f"The number of rows in the DataFrame is: {num_rows}")
```

```
     status gender      length  firstName level  lastName  registration  \
0       200      M   524.32934      Shlok  paid   Johnson  1.533735e+12
1       200      F   238.39302    Vianney  paid    Miller  1.537500e+12
2       200      F   140.35546       Vina  paid    Bailey  1.536415e+12
3       200      M   277.15873     Andres  paid     Foley  1.534387e+12
4       200      F  1121.25342    Aaliyah  paid   Ramirez  1.537381e+12
..      …    …          …         …   …       …           …
995     307      F         NaN     Alivia  paid  Williams  1.535955e+12
996     200      M   315.48036  Christian  free     Klein  1.536940e+12
997     200      M   248.73751  Kristofer  free     James  1.536879e+12
998     200      F    53.08036       Zoey  free   Gregory  1.533190e+12
999     200      F         NaN       Zoey  free   Gregory  1.533190e+12

      userId             ts       auth         page
0    1749042  1538352001000  Logged In     NextSong
1    1563081  1538352002000  Logged In     NextSong
2    1697168  1538352002000  Logged In     NextSong
3    1222580  1538352003000  Logged In     NextSong
4    1714398  1538352003000  Logged In     NextSong
..       …            …        …           …
995  1792538  1538352525000  Logged In    Thumbs Up
```

```
996  1291813  1538352526000  Logged In     NextSong
997  1929921  1538352526000  Logged In     NextSong
998  1839740  1538352527000  Logged In     NextSong
999  1839740  1538352527000  Logged In  Roll Advert

[1000 rows x 11 columns]

The number of rows in the DataFrame is: 26259199
```

[5]:
```
# Group by 'userId' and count the occurrences
user_counts = df.groupby('userId').size().reset_index(name='count')

# Display the result
display(user_counts.sort_values(by='count', ascending=False))
```

```
        userId   count
5936   1261737  778479
12534  1564221   13591
20802  1931933   12831
163    1006695   12372
7562   1336969   11858
…         …      …
6072   1267517       1
9698   1434698       1
17778  1793623       1
9129   1408726       1
15322  1689121       1

[22278 rows x 2 columns]
```

[6]:
```
print(pd.to_datetime(df['ts'].min(), unit='ms'))
print(pd.to_datetime(df['ts'].max(), unit='ms'))
```

```
2018-10-01 00:00:01
2018-12-01 00:00:02
```

[ ]: