Data Analysis in Polars and Pandas

Author: https://gist.github.com/koaning

Blog: https://calmcode.io/polars/introduction.html

Notebook: https://gist.github.com/koaning/5a0f3f27164859c42da5f20148ef3856

Dataset: https://www.kaggle.com/datasets/mylesoneill/warcraft-avatar-history?resource=download&select=wowah_data.csv

In [1]:
```
import polars as pl
```

In [3]:
```
pl.__version__

# '0.15.14'
```

Out[3]:
```
'0.15.14'
```

Let's do some stuff with a dataset!

Benchmark 1: Polars

In [4]:
```
datafile = "../data/kaggle/wowah_data.csv"   # 628 MB
```

In [5]:
```
%%time

df = pl.read_csv(datafile, parse_dates=False, n_threads=10)
df.columns = [c.replace(" ", "") for c in df.columns]
df = df.lazy()
```

```
Wall time: 890 ms
```

In [6]:
```
def set_types(dataf):
    return (dataf
            .with_columns([
                pl.col("timestamp").str.strptime(pl.Datetime, fmt="%m/%d/%y %H:%M:%S'
                pl.col("guild") != -1,
             ]))

def sessionize(dataf, threshold=1_000_000):
    return (dataf
             .sort(["char", "timestamp"])
             .with_columns([
                 (pl.col("timestamp").diff().cast(pl.Int64) > threshold).fill_null(Tru
                 (pl.col("char").diff() != 0).fill_null(True).alias("char_diff"),
             ])
             .with_columns([
                 (pl.col("ts_diff") | pl.col("char_diff")).alias("new_session_mark")
             ])
             .with_columns([
```

```
                pl.col("new_session_mark").cumsum().alias("session")
            ])
            .drop(['char_diff', 'ts_diff', 'new_session_mark']))

def add_features(dataf):
    return (dataf
            .with_columns([
                pl.lit(1).alias("one")
            ])
            .with_columns([
                pl.col("one").count().over("session").alias("session_length"),
                pl.col("session").n_unique().over("char").alias("n_sessions")
            ]))

def remove_bots(dataf, max_session_hours=24):
    n_rows = max_session_hours*6
    return (dataf
            .filter(pl.col("session_length").max().over("char") < n_rows))
```

In [7]:
```
df.collect().shape
```

Out[7]:
```
(10826734, 7)
```

In [8]:
```
%%time

(df
 .pipe(set_types)
 .pipe(sessionize)
 .pipe(add_features)
 .pipe(remove_bots)
 .collect())
```

```
Wall time: 3.84 s
```

Out[8]: shape: (10826734, 11)

char	level	race	charclass	zone	guild	timestamp	session	one	sessi
i64	i64	str	str	str	bool	datetime[µs]	u32	i32	
2	18	"Orc"	"Shaman"	"The Barrens"	true	2008-12-03 10:41:47	1	1	
7	54	"Orc"	"Hunter"	"Feralas"	false	2008-01-15 21:47:09	2	1	
7	54	"Orc"	"Hunter"	"Un'Goro Crater...	false	2008-01-15 21:56:54	3	1	
7	54	"Orc"	"Hunter"	"The Barrens"	false	2008-01-15 22:07:23	4	1	
7	54	"Orc"	"Hunter"	"Badlands"	false	2008-01-15 22:17:08	5	1	
7	54	"Orc"	"Hunter"	"Badlands"	false	2008-01-15 22:26:52	6	1	
7	54	"Orc"	"Hunter"	"Badlands"	false	2008-01-15 22:37:25	7	1	
7	54	"Orc"	"Hunter"	"Swamp of Sorro...	true	2008-01-15 22:47:10	8	1	
7	54	"Orc"	"Hunter"	"The Temple of ...	true	2008-01-15 22:56:53	9	1	
7	54	"Orc"	"Hunter"	"The Temple of ...	true	2008-01-15 23:07:25	10	1	
7	54	"Orc"	"Hunter"	"The Temple of ...	true	2008-01-15 23:17:09	11	1	
7	55	"Orc"	"Hunter"	"The Temple of ...	true	2008-01-15 23:26:53	12	1	
...	...	...	...	...	...	...	...	...	
90575	2	"Orc"	"Hunter"	"Orgrimmar"	false	2008-12-31 21:14:51	10823166	1	
90576	2	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:06:58	10823167	1	
90576	3	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:17:35	10823168	1	
90576	3	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:32:52	10823169	1	
90576	4	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:47:54	10823170	1	
90576	5	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 23:07:13	10823171	1	
90577	1	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:17:35	10823172	1	
90577	2	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:32:52	10823173	1	
90577	3	"Blood Elf"	"Warlock"	"Eversong Woods...	false	2008-12-31 22:47:54	10823174	1	
90578	1	"Blood Elf"	"Paladin"	"Eversong Woods...	false	2008-12-31 22:32:52	10823175	1	
90579	1	"Orc"	"Warrior"	"Durotar"	false	2008-12-31 22:44:45	10823176	1	
90580	1	"Tauren"	"Warrior"	"Mulgore"	false	2008-12-31 23:15:20	10823177	1	

wall-time: 3.84 sec shape: (10826734, 11)

Benchmark 2: Pandas

In [10]:
```
import pandas as pd
```

In [21]:
```
pd.__version__
```

Out[21]:
```
'1.4.4'
```

In [11]:
```
%%time

df = pd.read_csv(datafile)
df.columns = [c.replace(" ", "") for c in df.columns]

# Wall time: 6.68 s
```

```
Wall time: 6.68 s
```

In [17]:
```
df
```

Out[17]:

	char	level	race	charclass	zone	guild	timestamp
0	59425	1	Orc	Rogue	Orgrimmar	165	01/01/08 00:02:04
1	65494	9	Orc	Hunter	Durotar	-1	01/01/08 00:02:04
2	65325	14	Orc	Warrior	Ghostlands	-1	01/01/08 00:02:04
3	65490	18	Orc	Hunter	Ghostlands	-1	01/01/08 00:02:04
4	2288	60	Orc	Hunter	Hellfire Peninsula	-1	01/01/08 00:02:09
...	...	...	...	...	...	...	...
10826729	86766	80	Blood Elf	Death Knight	Halls of Lightning	101	12/31/08 23:50:18
10826730	86497	77	Blood Elf	Death Knight	The Storm Peaks	358	12/31/08 23:50:18
10826731	34893	80	Blood Elf	Death Knight	The Storm Peaks	189	12/31/08 23:50:18
10826732	86881	80	Blood Elf	Death Knight	Dragonblight	478	12/31/08 23:50:18
10826733	86457	80	Blood Elf	Death Knight	Dragonblight	204	12/31/08 23:50:18

10826734 rows × 7 columns

In [12]:
```
def set_types(dataf):
    return (dataf
            .assign(timestamp=lambda d: pd.to_datetime(d['timestamp'], format="%m/%d/%
                    guild=lambda d: d['guild'] != -1))

def sessionize(dataf, threshold=60*10):
    return (dataf
             .sort_values(["char", "timestamp"])
             .assign(ts_diff=lambda d: (d['timestamp'] - d['timestamp'].shift()).dt.se
                     char_diff=lambda d: (d['char'].diff() != 0),
                     new_session_mark=lambda d: d['ts_diff'] | d['char_diff'],
                     session=lambda d: d['new_session_mark'].fillna(0).cumsum())
             .drop(columns=['char_diff', 'ts_diff', 'new_session_mark']))

def add_features(dataf):
    return (dataf
              .assign(session_length=lambda d: d.groupby('session')['char'].transform(
              .assign(n_sessions=lambda d: d.groupby('char')['session'].transform(lamb

def remove_bots(dataf, max_session_hours=24):
    n_rows = max_session_hours*6
    return (dataf
            .assign(max_sess_len=lambda d: d.groupby('char')['session_length'].transfo
            .loc[lambda d: d["max_sess_len"] < n_rows]
            .drop(columns=["max_sess_len"]))
```

In [13]:
```
%%time

dataf = df.pipe(set_types).pipe(sessionize)
```

```
Wall time: 16.9 s
```

Wall time: 16.9 s

In [14]:
```
%%time

final = dataf.pipe(add_features).pipe(remove_bots)
```

```
Wall time: 8min 30s
```

Wall time: 8min 30s

The Results?

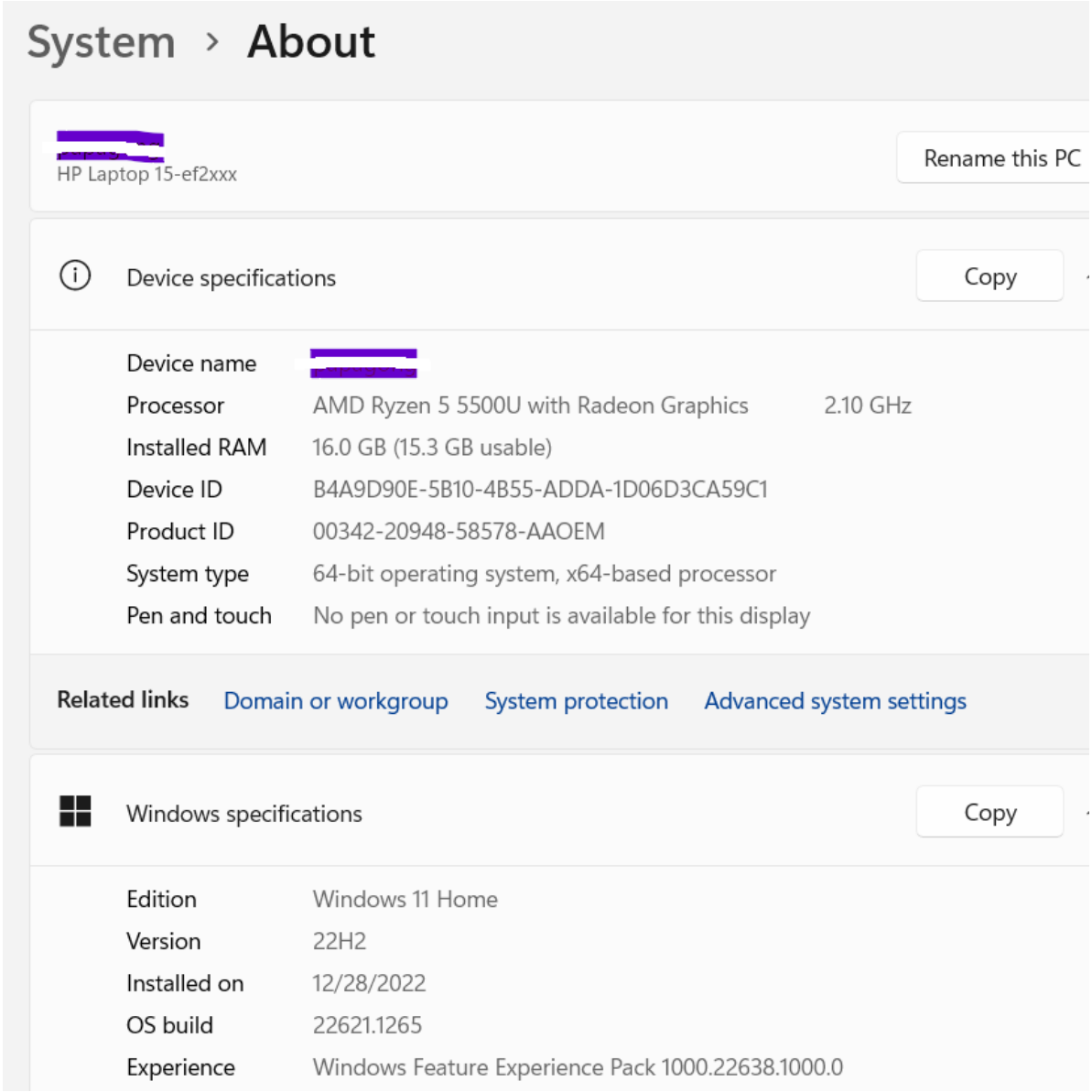
- polars `3.84 sec`
- pandas `8m 47s`

It's not a perfect benchmark, and it depends a bit on how on measures ... but a rough speedup factor is:

In [20]:
```
polars_over_pandas_speedup_factor = (8*60+47)/3.84
print(f"Polars over Pandas speedup factor: {polars_over_pandas_speedup_factor:.2f}, Ho
```

```
Polars over Pandas speedup factor: 137.24, Hooray!!!
```

System info


System › About
[redacted]
HP Laptop 15-ef2xxx
Rename this PC
Device specifications
Copy
Device name [redacted]
Processor AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz
Installed RAM 16.0 GB (15.3 GB usable)
Device ID B4A9D90E-5B10-4B55-ADDA-1D06D3CA59C1
Product ID 00342-20948-58578-AAOEM
System type 64-bit operating system, x64-based processor
Pen and touch No pen or touch input is available for this display
Related links Domain or workgroup System protection Advanced system settings
Windows specifications
Copy
Edition Windows 11 Home
Version 22H2
Installed on 12/28/2022
OS build 22621.1265
Experience Windows Feature Experience Pack 1000.22638.1000.0

In [ ]: