Introduction about dataset

The Dataset I used here is collected from Kaggle . it is an ecommerce dataset that contains data about customer of a supershop with fields :

- InvoiceNo
- StockCode
- Description
- Quantity
- InvoiceDate
- UnitPrice
- CustomerID
- Country

This dataset contains about 541909 transactions from the date of january 2010 to december 2011.

For more details: Dataset

Working Process:

Used Python Libraries :

1. Pandas
2. Pandasql
3. Fastparquet
4. Os

Step-1:

Read data from data.csv file and load them in a pandas dataframe.

```
df = pd.read_csv('./data.csv', encoding = 'unicode_escape')
```

Step-2:

Split Invoicedate field into separate date and time column.

```
df[["date","time"]] = df.InvoiceDate.str.split(' ',n=1,expand=True)
```

Step-3:

Reform the date format

```
df[['date']] = df.date.str.replace('/','-')
```

Step-4:

Pick all the unique date from the dataframe

```
all_distinct_date = "select distinct(date) from df"
all_distinct_date_df = ps.sqldf(all_distinct_date, locals())
```

Step-5:

Make a parent Directory to keep all the sub folder

```
os.mkdir('result')
parent_dir = 'result/'
```

Step-6:

Loop over all the unique date field

```
for index,item in all_distinct_date_df.iterrows():
```

Step-7:

Make Directory by unique date

```
path = os.path.join(parent_dir, date_dir_name)
os.mkdir(path)
```

Step-8:

Get all Transaction from a unique date

```
all_ochurance_in_a_day = 'select * from df where date = "'+date_dir_name+'"'
all_ochurance_in_a_day_df = ps.sqldf(all_ochurance_in_a_day, globals())
```

Step-9:

Get all distinct transection hour from step-8

```
all_distinct_hour_in_a_day = 'select distinct(time) from all_ochurance_in_a_day_df'
all_distinct_hour_in_a_day_df = ps.sqldf(all_distinct_hour_in_a_day, globals())
```

Step-10:

Loop over all the distinct transection hour

```
for index,val in all_distinct_hour_in_a_day_df.iterrows():
```

Step-11:

Make directory by time name

```
time_dir_path = os.path.join(inside_date_folder,hour_dir_name)
    os.mkdir(time_dir_path)
```

Step-12:

Get all transaction in a unique hour

```
    all_occurance_in_a_hour = "select * from all_ochurance_in_a_day_df where time =
'"+hour_dir_name+"'"
all_occurance_in_a_hour_df = ps.sqldf(all_occurance_in_a_hour, globals())
```

Step-13:

Make file name with parquet format

```
    put_per = time_dir_path+'/'+hour_dir_name+'.parquet.gzip'
```

Step-14:

Make a parquet file with a unique transaction and put them inside the path.

```
all_occurance_in_a_hour_df.to_parquet(put_per,compression='gzip'
)
```

Run the Docker File:

Make image :

docker build -t ImageName .

Run the image:

docker run --rm -it -v {PWD}:/app Imagename ker build -t park .

____END_____