In [1]:

```
import pandas as pd
import os
from tqdm import tqdm_notebook as tqdm
import dask.dataframe as dd
from dask.diagnostics import ProgressBar
ProgressBar().register()
import multiprocessing
nCPU = multiprocessing.cpu_count()
from datetime import datetime, timedelta
```

```
/Users/koheiyamamoto/.pyenv/versions/anaconda3-5.3.1/lib/python3.7/site-packages/d
ask/dataframe/utils.py:13: FutureWarning: pandas.util.testing is deprecated. Use t
he functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

In [2]:

```
# Get paths to files under the pwd, returned to get_df()
def get_paths(c_path, extension, yyyymmdd, limitation_keyword):
  l = []
  for pathname, dirnames, filenames in os.walk(c_path):
      for filename in filenames:
          if yyyymmdd in filename.split('.')[0] and filename.split('.')[
              formal_path = os.path.join(pathname, filename)
              if limitation_keyword == '': # when no keyword is specifie
                  l.append(formal_path)
              else:
                  if limitation_keyword in formal_path.split('/'): # whe
                      l.append(formal_path)
  return l

def get_date(x):
    t = datetime.strptime(x, "%Y-%m-%d %H:%M:%S")
    date = str(t.year) + '-' + str("%02d" % t.month) + '-' + str("%02d" %
    if t.year != 2017: # 2016 data is by accident included
        return None
    else:
        return date

def get_timeperiod(x):
    t = datetime.strptime(x, "%Y-%m-%d %H:%M:%S")
    if t.hour in [22, 23, 0, 1, 2, 3]:
        tp = 'latenight'
    elif t.hour in [4, 5, 6, 7, 8, 9]:
        tp = 'morning'
    elif t.hour in [10, 11, 12, 13, 14, 15]:
        tp = 'midday'
    else:
        tp = 'evening'
    return tp
```

In [3]:

```
def aggregate_taxidata(datadir, dataname):
    df_out = pd.DataFrame(index=[], columns=['DOLocationID', 'Date', 'Ti
    for i in tqdm(get_paths('./data/' + datadir, 'csv', dataname, '')):
        df = pd.read_csv(i, dtype=str)

        if datadir == 'yellow':
            dropField = ['VendorID', 'tpep_pickup_datetime', 'passenger_
            df = df.drop(dropField, axis='columns')
        elif datadir == 'green':
            dropField = ['VendorID', 'lpep_pickup_datetime', 'store_and_
            df = df.drop(dropField, axis='columns')
            df.rename(inplace=True, columns={"lpep_dropoff_datetime": "t
        elif datadir == 'fhv':
            dropField = ['Dispatching_base_num', 'Pickup_DateTime', 'PUl
            df = df.drop(dropField, axis='columns')
            df = df.dropna(subset=['DropOff_datetime'])
            df = df.dropna(subset=['DOlocationID'])
            df.rename(inplace=True, columns={"DropOff_datetime": "tpep_d

        df = dd.from_pandas(df, npartitions=nCPU)
        meta = df['tpep_dropoff_datetime'].head(1).apply(lambda x: get_d
        res = df['tpep_dropoff_datetime'].apply(lambda x: get_date(x), m
        df['Date'] = res.compute(scheduler='processes')
        meta = df['tpep_dropoff_datetime'].head(1).apply(lambda x: get_t
        res = df['tpep_dropoff_datetime'].apply(lambda x: get_timeperiod
        df['Time_Period'] = res.compute(scheduler='processes')
        df = df.compute()

        df = df.dropna(subset=['Date'])
        df = df.drop(['tpep_dropoff_datetime'], axis='columns')
        df_out = pd.concat([df_out, df])

    df_out.reset_index(inplace=True, drop=True)
    df_agg_out = pd.DataFrame(index=[], columns=['Date','DOLocationID','
    for i in tqdm(df_out.Date.unique().tolist()):
        df_tmp = df_out.query('Date == @i')

        tmp = df_tmp.groupby('DOLocationID').Time_Period.value_counts().
        tmp['Date'] = i
        tmp.reset_index(inplace=True)
```

```
        tmp = tmp[['Date','DOLocationID','Time_Period','Volume']]

        df_agg_out = pd.concat([df_agg_out, tmp])
    df_agg_out.rename(inplace=True, columns={"Date": "Date", "DOLocation
    df_agg_out.to_csv(dataname + '_out.csv', index=False)
```

In [4]:

```
aggregate_taxidata('yellow', 'yellow_tripdata_2017')
```

100% 12/12 [2:05:34<00:00, 619.75s/it]

```
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  7.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8min 31.4s
[########################################] | 100% Completed |  1.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 12.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  9min  3.6s
[########################################] | 100% Completed |  1.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  4.7s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8min 21.2s
[########################################] | 100% Completed |  1.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 12.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  9min 14.0s
[########################################] | 100% Completed |  1.4s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 20.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  9min  7.3s
[########################################] | 100% Completed |  1.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  2.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7min 43.1s
[########################################] | 100% Completed |  1.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 12.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  9min  2.5s
[########################################] | 100% Completed |  1.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  7.7s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8min 24.0s
[########################################] | 100% Completed |  1.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 12.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8min 51.3s
[########################################] | 100% Completed |  1.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 12.6s
[########################################] | 100% Completed |  0.1s
```

```
[########################################] | 100% Completed |  9min  1.2s
[########################################] | 100% Completed |  1.3s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  6.8s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8min 10.6s
[########################################] | 100% Completed |  1.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min  4.0s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7min 43.0s
[########################################] | 100% Completed |  1.1s
```

100% 365/365 [21:02<00:00,

3.22s/it]

In [5]:

```
aggregate_taxidata('green', 'green_tripdata_2017')
```

100% 12/12 [12:14<00:00, 61.55s/it]

```
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 44.1s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.6s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 44.4s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7.8s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 53.6s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7.5s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 51.4s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8.6s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 58.4s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.8s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 46.1s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 49.2s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.7s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 45.8s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  8.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 54.5s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.8s
[########################################] | 100% Completed |  0.1s
```

```
[########################################] | 100% Completed | 47.1s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  6.5s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 44.4s
[########################################] | 100% Completed |  0.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  7.7s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 53.5s
[########################################] | 100% Completed |  0.2s
```

100% 365/365 [01:50<00:00,

3.44it/s]

In [6]:

```
aggregate_taxidata('fhv', 'fhv_tripdata_2017')
```

100% 12/12 [2:04:22<00:00, 948.25s/it]

```
[########################################] | 100% Completed |  0.1s

/Users/koheiyamamoto/.pyenv/versions/anaconda3-5.3.1/lib/python3.7/site-packages/dask/dataframe/core.py:4347: UserWarning: Insufficient elements for `head`. 1 elements requested, only 0 elements available. Try passing larger `npartitions` to `head`.
  warnings.warn(msg.format(n, len(r)))

[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  2min 18.4s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 15min  9.1s
[########################################] | 100% Completed |  3.0s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s

/Users/koheiyamamoto/.pyenv/versions/anaconda3-5.3.1/lib/python3.7/site-packages/dask/dataframe/core.py:4347: UserWarning: Insufficient elements for `head`. 1 elements requested, only 0 elements available. Try passing larger `npartitions` to `head`.
  warnings.warn(msg.format(n, len(r)))

[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  0.1s
```

```
[########################################] | 100% Completed |  2min 15.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 15min  1.7s
[########################################] | 100% Completed |  2.9s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  1min 39.9s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 11min 47.3s
[########################################] | 100% Completed |  1.8s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  2min 32.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 16min 53.3s
[########################################] | 100% Completed |  3.2s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  2min  0.9s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 13min 21.4s
[########################################] | 100% Completed |  2.6s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  2min  9.1s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 14min  5.2s
[########################################] | 100% Completed |  2.7s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed |  2min  2.5s
[########################################] | 100% Completed |  0.1s
[########################################] | 100% Completed | 13min 34.6s
[########################################] | 100% Completed |  2.6s
```

100% 216/216 [11:16<00:00,

3.63s/it]

In [ ]: