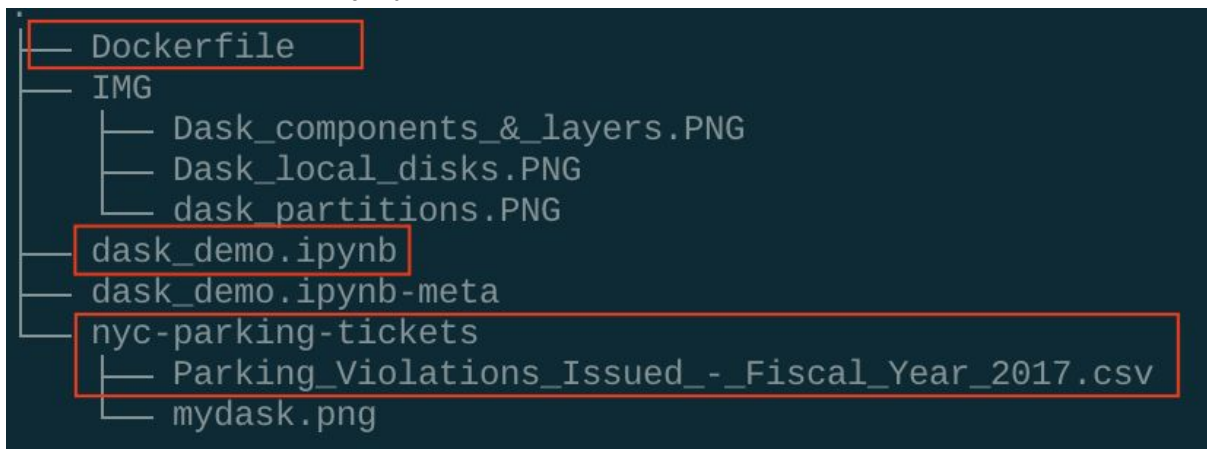


## Local Environment

1. **Create a folder**
2. **Download from this website and put inside the folder**  
<https://harvard-iacs.github.io/2020F-AC295/lectures/lecture4/>
  - a. Dockerfile
  - b. dask\_demo jupyter notebook



red boxes indicate the folders & files you need

3. **Download the data**  
<https://www.kaggle.com/new-york-city/nyc-parking-tickets>
4. **Install dask: [Tutorial](#)**  
conda install dask  
or  
python -m pip install "dask[complete]"
5. **Install graphviz** (you probably have it already)
  - a. brew install graphviz (for Mac users)  
<https://graphviz.org/download/>
  - b. pip3 install graphviz  
(both commands have to be run - need graphviz on your system and for python)
6. **Provide the path to the data in the jupyter notebook**

## 2.1.1 Set up environment and working directory

```
[1]: # import libraries
import sys
import os

## import dask libraries
import dask.dataframe as dd
from dask.diagnostics import ProgressBar

# import libraries
import pandas as pd

[3]: # assign working directory [CHANGE THIS]. It can not fit in github so I have it locally. Download files
os.chdir('/Users/...your path here... /daskdemo/notebook/nyc-parking-tickets')
cwd = os.getcwd()

# print
print('<environment path>', sys.executable)
print('<current working directory>', cwd)

<environment path> /Users/haibui/.pyenv/versions/3.7.7/bin/python3.7
```

## 7. Run the notebook

## Dockerize it

1. Create a folder
2. Download from this website and put inside the folder  
<https://harvard-iacs.github.io/2020F-AC295/lectures/lecture4/>
  - a. Dockerfile
  - b. dask\_demo jupyter notebook

```
├── Dockerfile
├── IMG
│   ├── Dask_components_&_layers.PNG
│   ├── Dask_local_disks.PNG
│   └── dask_partitions.PNG
├── dask_demo.ipynb
├── dask_demo.ipynb-meta
├── nyc-parking-tickets
│   └── Parking_Violations_Issued_-_Fiscal_Year_2017.csv
└── mydask.png
```

red boxes indicate the folders & files you need

3. Download the data  
<https://www.kaggle.com/new-york-city/nyc-parking-tickets>
4. Put the data inside your folder (see tree structure above)
  - a. If you want to save time delete everything except "Parking\_Violations\_Issued\_-\_Fiscal\_Year\_2017.csv" in the data folder

### In Terminal (on your host OS)

5. `docker build -t daskdemo .`

```
haibui@Hais-MBP ~ -/00_MIT_Harvard_CS_DS/harvard_data_science/daskdemo/notebook [master] docker build -t daskdemo .
Sending build context to Docker daemon  5.861MB
Step 1/9 : FROM ubuntu:latest
--> 4e2eef94cd6b
Step 2/9 : RUN apt update -y
--> Using cache
--> d2142ba4d60b
Step 3/9 : RUN apt install -y python3-pip
--> Using cache
--> 6e17b0d1bac2
Step 4/9 : RUN python3 -m pip install jupyterlab
--> Using cache
--> 8d2d8f583fc4
Step 5/9 : RUN python3 -m pip install "dask[complete]"
--> Using cache
--> deeebe8cd6e3
Step 6/9 : RUN python3 -m pip install graphviz
--> Using cache
--> b42c36779b3a
Step 7/9 : COPY . .
--> 749044fdd40b
Step 8/9 : EXPOSE 9999
--> Running in 7c33b7730c23
Removing intermediate container 7c33b7730c23
--> 7aad8521abb4
Step 9/9 : CMD ["/bin/sh"]
--> Running in b0cddb12021d
Removing intermediate container b0cddb12021d
--> d13e1de1015b
Successfully built d13e1de1015b
Successfully tagged daskdemo:latest
```

6. `sudo docker run -p 9999:9999 -ti daskdemo`
  - a. input password (of your mac)
  - b. this will get you inside the container

```
halbui@Hais-MBP ~ /00 MIT Harvard CS DS/harvard data science/daskdemo/notebook | master | sudo docker run -p 9999:9999 -ti daskdemo
# ls
Dockerfile IMG bin boot dask_demo.ipynb dask_demo.ipynb-meta dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
# apt-get install graphviz
Reading package lists... Done
Building dependency tree
```

## Inside the container

7. apt-get install graphviz
  - a. insert: y, 2 and 31 (or whatever is right for your location)

```
# apt-get install graphviz
Reading package lists... Done
Building dependency tree
Reading state information... Done
```

8. jupyter lab --ip='0.0.0.0' --port=9999 --no-browser --allow-root  
(this will run the jupyter lab)
  - a. Copy the link provided in the terminal to browser

```
# jupyter lab --ip='0.0.0.0' --port=9999 --no-browser --allow-root
[I 06:16:12.734 LabApp] Writing notebook server cookie secret to /root/.local/share/jupyter/runtime/notebook_cookie_secret
[I 06:16:12.948 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.8/dist-packages/jupyterlab
[I 06:16:12.949 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab
[I 06:16:12.954 LabApp] Serving notebooks from local directory: /
[I 06:16:12.954 LabApp] Jupyter Notebook 6.1.4 is running at:
[I 06:16:12.955 LabApp] http://d9b5b7f81d73:9999/?token=c3a64b4e389a6d368c3024858c1a13fa34ff1e79926dda77
[I 06:16:12.956 LabApp] or http://127.0.0.1:9999/?token=c3a64b4e389a6d368c3024858c1a13fa34ff1e79926dda77
[I 06:16:12.956 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 06:16:12.960 LabApp]

To access the notebook, open this file in a browser:
file:///root/.local/share/jupyter/runtime/nbserver-663-open.html
Or copy and paste one of these URLs:
http://d9b5b7f81d73:9999/?token=c3a64b4e389a6d368c3024858c1a13fa34ff1e79926dda77
or http://127.0.0.1:9999/?token=c3a64b4e389a6d368c3024858c1a13fa34ff1e79926dda77
```

9. Run notebook