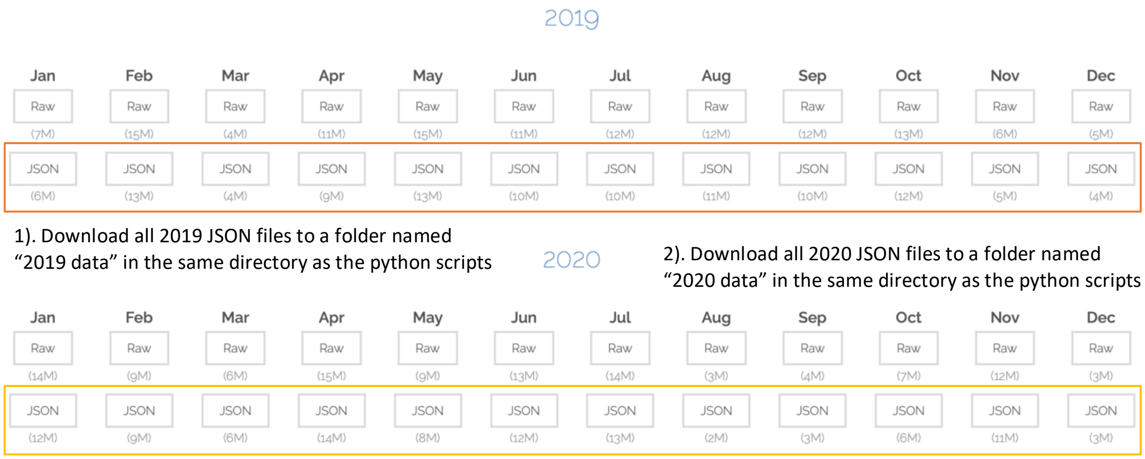
Instructions for running the codes

Note:
You are advised to skip steps 1-4, which include downloading of 36 GB of data and usage of pySpark for data cleaning. The cleaned data files are provided in the "processed data" folder for you to directly start from step 5.

1. Install pySpark. You may refer to https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes

2. Download AIS data from https://data.liancheng.science/ais_logs.html



2019

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Raw (7M)	Raw (15M)	Raw (4M)	Raw (11M)	Raw (15M)	Raw (11M)	Raw (12M)	Raw (12M)	Raw (12M)	Raw (13M)	Raw (6M)	Raw (5M)
JSON (6M)	JSON (13M)	JSON (4M)	JSON (9M)	JSON (13M)	JSON (10M)	JSON (10M)	JSON (11M)	JSON (10M)	JSON (12M)	JSON (5M)	JSON (4M)

1). Download all 2019 JSON files to a folder named "2019 data" in the same directory as the python scripts

2020

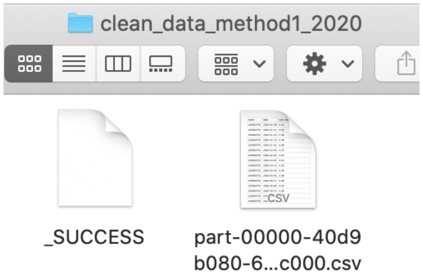
2). Download all 2020 JSON files to a folder named "2020 data" in the same directory as the python scripts

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Raw (14M)	Raw (9M)	Raw (6M)	Raw (15M)	Raw (9M)	Raw (13M)	Raw (14M)	Raw (3M)	Raw (4M)	Raw (7M)	Raw (12M)	Raw (3M)
JSON (12M)	JSON (9M)	JSON (6M)	JSON (14M)	JSON (8M)	JSON (12M)	JSON (13M)	JSON (2M)	JSON (3M)	JSON (6M)	JSON (11M)	JSON (3M)

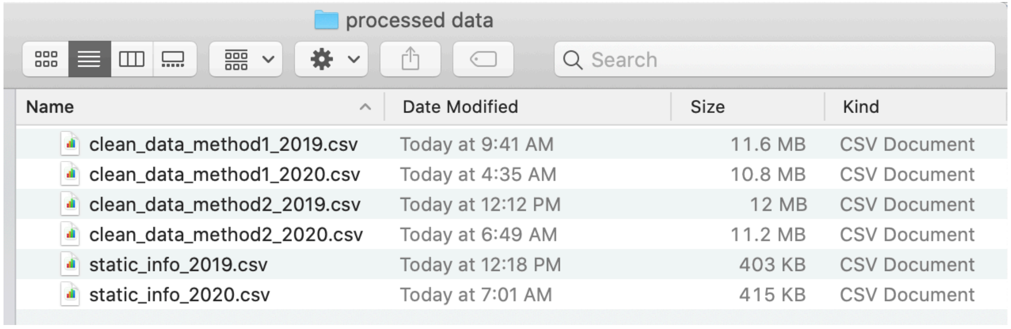
3. Run "AIS Data Cleaning.ipynb". Note that the whole script may take up to 20 hours to finish running.

4. There are 6 output **folders** generated from step 3.
    a. clean_data_method1_2019, clean_data_method1_2020
    b. clean_data_method2_2019, clean_data_method2_2020
    c. static_info_2019, static_info_2020

    We need to manually rename the csv file within each folder as its folder name. We shall then take out the csv file and delete the folder. For instance, we need to rename the "part-00000-……" file in the screenshot below as "clean_data_method1_2020.csv".



After doing so for all 6 files, we place them into a folder named “processed data” in the same directory as the python scripts, as shown in the screenshot below.



Name	Date Modified	Size	Kind
clean_data_method1_2019.csv	Today at 9:41 AM	11.6 MB	CSV Document
clean_data_method1_2020.csv	Today at 4:35 AM	10.8 MB	CSV Document
clean_data_method2_2019.csv	Today at 12:12 PM	12 MB	CSV Document
clean_data_method2_2020.csv	Today at 6:49 AM	11.2 MB	CSV Document
static_info_2019.csv	Today at 12:18 PM	403 KB	CSV Document
static_info_2020.csv	Today at 7:01 AM	415 KB	CSV Document

--data cleaning finished--

5. Run “Data Analysis Notebook.ipynb”.
   Note that the section “**Pull vessel particulars from SG-MDH using API**” is commented out and the result file “vessel_info.csv” is provided in “SG MDH data” folder. If you wish to generate this file by yourself, just uncomment this section and it takes around 30 minutes to finish running. In addition, a country code reference table named “COUNTRIES_REF_JSON.json” is pre-downloaded and contained in the same folder.