Compression of data and vessel traffic analysis surrounding COVID-19 years

Andrew Mueller, Peyton ...

Problem

- On average, around 60,000 vessels will arrive at a U.S. port annually
- Vessel movement can be affected by a multitude of factors, including trade policy, war, and global pandemics
- Our dataset is large, on the order of hundreds of GB

Goals

- Analyze the impact of COVID-19 on vessel traffic at West Coast ports (Los Angeles and Long Beach) by comparing 2019 and 2020 data.
- Examine disruptions to U.S. port traffic caused by pandemics, weather, and trade policy changes.
- Introduce a lossless compression algorithm that reduces maritime dataset size.

Dataset

- Data comes from marinecadastre.gov
- Stored and distributed in (CSV) format, filtered to one minute, and formatted in daily files for all U.S. coastal waters
- Average size of data files: 740 MB
- Data starts from 2009 - Current
- Each broadcast is separated on each line

MMSI	BaseDateTime	LAT	LON	SOG	COG	Heading	VesselName	IMO	CallSign	VesselType	Status	Length	Width	Draft	Cargo	TransceiverClass
367060130	2019-01-01 0:00:	29.68313	-91.1679	5.6	282.4	511					0					A
368029640	2019-01-01 0:00:	40.68142	-74.01218	0	360	511					0					A
367141210	2019-01-01 0:00:	39.8976	-75.13971	0.1	198.8	511					0					A

Dataset fields

	Name	Description	Example	Units	Resolution	Type	Size
1	MMSI	Maritime Mobile Service Identity value	477220100			Text	9
2	BaseDateTime	Full UTC date and time	2017-02-01T20:05:07		YYYY-MM-DD:HH-MM-SS	DateTime	
3	LAT	Latitude	42.35137	decimal degrees	XX.XXXXX	Double	8
4	LON	Longitude	-71.04182	decimal degrees	XXX.XXXXX	Double	8
5	SOG	Speed Over Ground	5.9	knots	XXX.X	Float	4
6	COG	Course Over Ground	47.5	degrees	XXX.X	Float	4
7	Heading	True heading angle	45.1	degrees	XXX.X	Float	4
8	VesselName	Name as shown on the station radio license	OOCL Malaysia			Text	32
9	IMO	International Maritime Organization Vessel number	IMO9627980			Text	7
10	CallSign	Call sign as assigned by FCC	VRME7			Text	8
11	VesselType	Vessel type as defined in NAIS specifications	70			Integer	short
12	Status	Navigation status as defined by the COLREGS	3			Integer	short
13	Length	Length of vessel (see NAIS specifications)	71.0	meters	XXX.X	Float	4
14	Width	Width of vessel (see NAIS specifications)	12.0	meters	XXX.X	Float	4
15	Draft	Draft depth of vessel (see NAIS specifications)	3.5	meters	XXX.X	Float	4
16	Cargo	Cargo type (see NAIS specification and codes)	70			Text	4
17	TransceiverClass	Class of AIS transceiver	A			Text	2

Compression - Goals

- One aim is to compress data to a more manageable size for scientists and data analysts to work with
- Using some simple observations of the data, the field types, and the resolution of certain fields, we were able to get about 5.7 compression ratio
- In compressing data, we also aim to get a speed up of reading in the data
- We aim for the compression to be lossless

Compression - Approach

- We designed our algorithm based on key insights into data types
- Using bit shifting, and understanding of primitive types, we can condense fields without losing information
- DateTime field can be broken into 2 components. The date component can be stored once per file, where as the time field needs to be stored for each unique broadcast point.
    - We are able to fit each component into 3 bytes with some bit shifting

Compression - Assumptions

- We assume the following byte sizes for each type we utilize:
  - long long will be 8 bytes,
  - Integer will be 4 bytes
  - Short will be 2 bytes
  - Float is 4 bytes.
- We assume that the only fields that will vary will be fields; (2, 3, 4, 5, 6, 7). All other fields should remain the same.
  - This assumption has not been confirmed, further analysis will need to be done
  - We assume this because these are the fields that pertain to the vessels location, and positioning at each time point that is collected. All other fields can be considered metadata about the vessel

Compression - Double to int example

- Latitude and longitude both have a set resolution
- If we use some bit logic, we can extract each digit into 4 bits
- Let’s encode 49.43831 to an integer value

Half byte number	7	6	5	4	3	2	1	0
Portion	U	U	U	I	I	I	I	I
Example	0	4	9	4	3	8	3	1

- For longitude and latitude, we save 4 bytes each

Compression - Other float fields

- For all other float fields, we can follow very similar logic that we used for the latitude and longitude
- Because the resolution of all other float fields is 4 digits, we only need a short for each of the fields

Compression Text fields

- Little can be done with text fields without losing information.
- We encode the size of each field
  - If size is zero, we don't need to write further on this field. Otherwise, we write the field
  - Helps in speeding up decompression since we know ahead of time the size of the field
-

4	T	e	s	t
Char	Char	Char	Char	Char

Compression - reducing data to bytes

MMSI	Date	Name length	Name	IMO	Callsign len	Callsign
4	3	1	l(name)	4	1	l(callsign)

Vessel Type	Status	Length	Width	Draft	Cargo len	Cargo
2	1	4	4	4	1	l(cargo)

TC len	Transciever Class	TS length	time	lat	lon	SOG	COG	Heading
1	l(transciever Class)	4	3	4	4	2	2	2

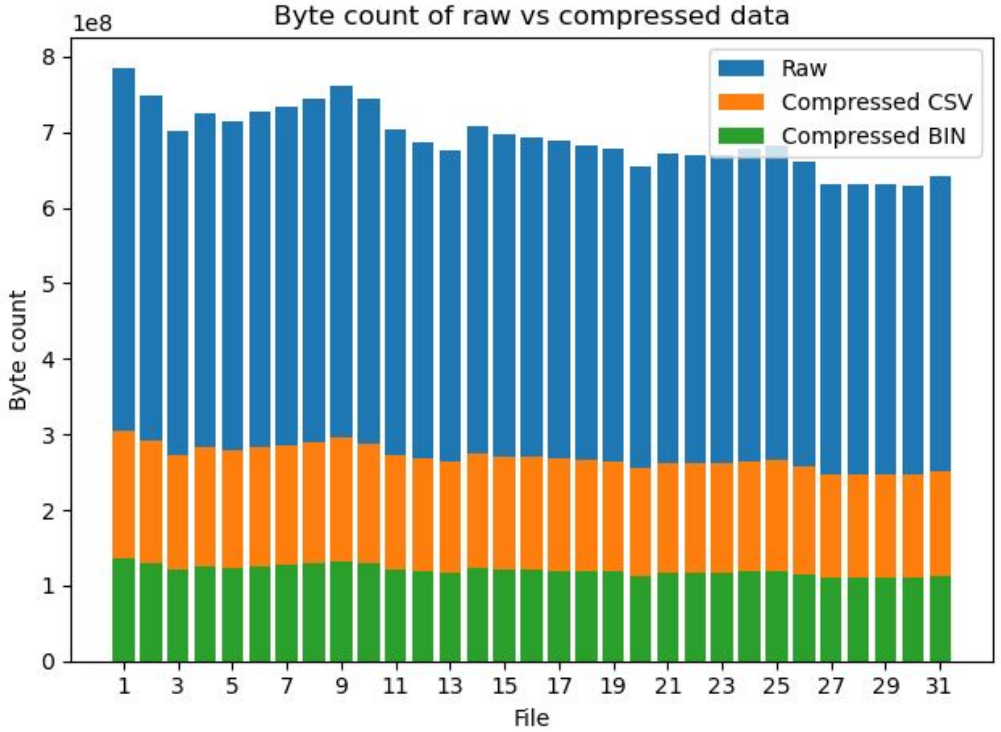
- We define the byte sizes for each field
- The last row is repeated for each broadcast point
  - Fields 2 -7
- l(FIELD) → Length of field in bytes

Compression - Main loop

- Using the MMSI as a key, we can create an unordered map of vessels with a class object to contain all relevant information about each vessel
-

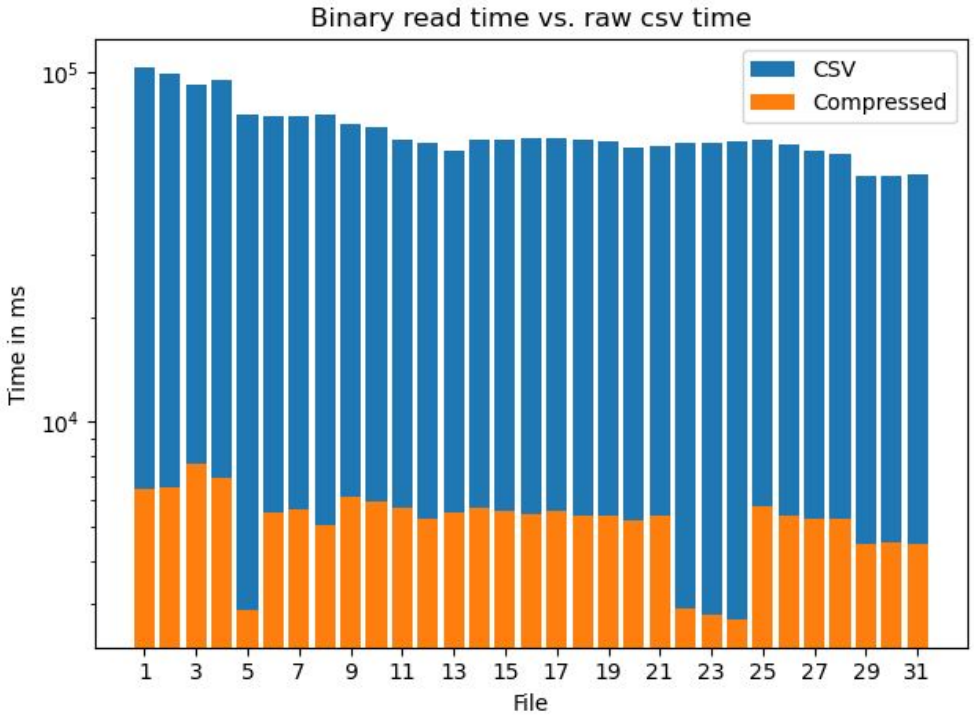
```
unordered_map<MMSI, ship*> vessels;

for (line in file)
    f1 = line{0, 1, 8 - 16}];
    f2 = line{1 - 7};
    if (f1[0] not in vessels)
        insert <vessels[MMSI], ship>;
        vessels[MMSI]->insert(f1);
    vessels[MMSI]->insert(f2);
```

Compression results : File size comparison



- Data for files for the month of January 2019
- Achieve average 5.7 compression ratio for BIN
- Achieve average 2.4 for CSV

Compression results : Read times



- Able to achieve a speedup of reading files in of approx 13x
- Allows for large number of files to be read in much faster.

Compression Results : statistics

Statistic	Speed up	Space savings %	Compression ratio
Average	13.65	82.63	5.76
Min	10.96	82.56	5.733
Max	26.09	82.72	5.79

- High compression ratio for lossless compression
- Generally high speed up in reading files
- Enormous space savings per file

Verification from raw to bin

- Verification steps
  - Read in raw csv file
  - Compress information to a binary file
  - Read in compressed binary file
  - Compare the information for both. If they are the same, we consider the compression lossless
- In our verification, we did not find that any of the data was different / lost for each of the vessel entries.
- Note: We would like to do further analysis on how much information we lose.

Future work

- Look at using map-reduce to help in speed up analysis
  - This would be a good approach since the files are quite large
- Further verify how much information is lost, if any
- Introduce some meta data in order to allow for multithreading reading of the compressed binary data, and analyze the speedup compared to sequential reading
-

Sources

- **Marine Cadastre (2016):** Overview of vessel traffic data. MarineCadastre.gov (Accessed Nov. 21, 2024).
- **Bureau of Transportation Statistics (2017):** Maritime trade and transportation statistics. BTS.gov.
- **USITC (2020):** Analysis of COVID-19's impact on freight transportation and U.S. merchandise imports. USITC.gov.
- **Marine Cadastre (AIS Data):** Automatic Identification System (AIS) data resource. MarineCadastre.gov (Accessed Nov. 28, 2024).
- **Marine Cadastre (Data Dictionary):** AIS data dictionary resource. NOAA.gov.