Scalable Data Analysis

Project Proposal

Standard Metadata:

Title: Vessels Traffic Data
Author: Ravi Shankar
Date: 11/20/2017

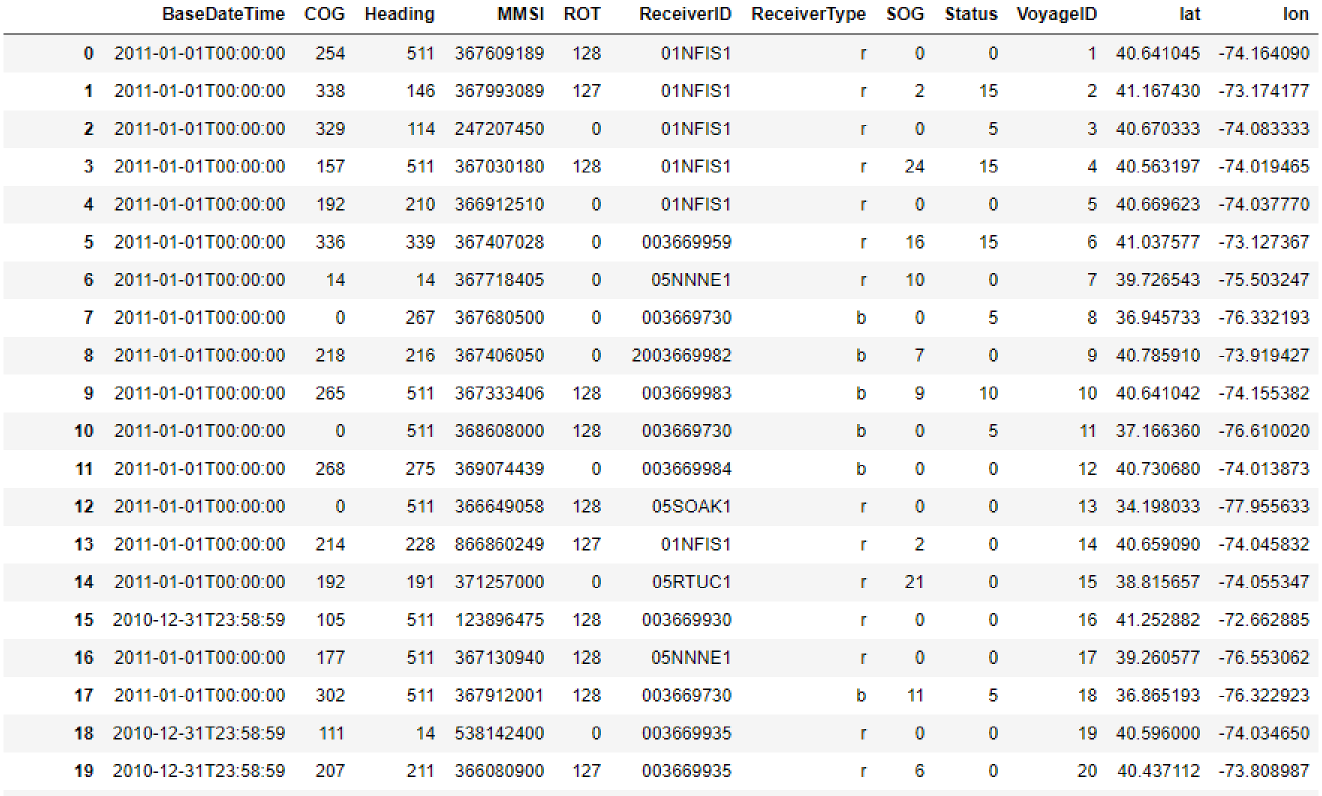
Introduction: I am working on the AIS Vessels Traffic Data, there is a large amount of trajectory data available, this data is taken from their webpage Marine Cadastre – Vessels traffic data, available here on the following webpage

https://marinecadastre.gov/ais/

This data is divided into a hierarchy of Year, and sub hierarchy of Zones and by Month, even available the full years data to download. Data is divided into Vessels information, broadcast data and the Voyages data,

Goals: There is so much work is already done before on this AIS Vessels traffic data. My goal for this project is to figure out the stops of the ships trajectory data, and I can segment those trajectories based on those stops, and can find how long that ships stayed at the place.

Following is the more detail about the Broadcast Data points.



	BaseDateTime	COG	Heading	MMSI	ROT	ReceiverID	ReceiverType	SOG	Status	VoyageID	lat	lon
0	2011-01-01T00:00:00	254	511	367609189	128	01NFIS1	r	0	0	1	40.641045	-74.164090
1	2011-01-01T00:00:00	338	146	367993089	127	01NFIS1	r	2	15	2	41.167430	-73.174177
2	2011-01-01T00:00:00	329	114	247207450	0	01NFIS1	r	0	5	3	40.670333	-74.083333
3	2011-01-01T00:00:00	157	511	367030180	128	01NFIS1	r	24	15	4	40.563197	-74.019465
4	2011-01-01T00:00:00	192	210	366912510	0	01NFIS1	r	0	0	5	40.669623	-74.037770
5	2011-01-01T00:00:00	336	339	367407028	0	003669959	r	16	15	6	41.037577	-73.127367
6	2011-01-01T00:00:00	14	14	367718405	0	05NNNE1	r	10	0	7	39.726543	-75.503247
7	2011-01-01T00:00:00	0	267	367680500	0	003669730	b	0	5	8	36.945733	-76.332193
8	2011-01-01T00:00:00	218	216	367406050	0	2003669982	b	7	0	9	40.785910	-73.919427
9	2011-01-01T00:00:00	265	511	367333406	128	003669983	b	9	10	10	40.641042	-74.155382
10	2011-01-01T00:00:00	0	511	368608000	128	003669730	b	0	5	11	37.166360	-76.610020
11	2011-01-01T00:00:00	268	275	369074439	0	003669984	b	0	0	12	40.730680	-74.013873
12	2011-01-01T00:00:00	0	511	366649058	128	05SOAK1	r	0	0	13	34.198033	-77.955633
13	2011-01-01T00:00:00	214	228	866860249	127	01NFIS1	r	2	0	14	40.659090	-74.045832
14	2011-01-01T00:00:00	192	191	371257000	0	05RTUC1	r	21	0	15	38.815657	-74.055347
15	2010-12-31T23:58:59	105	511	123896475	128	003669930	r	0	0	16	41.252882	-72.662885
16	2011-01-01T00:00:00	177	511	367130940	128	05NNNE1	r	0	0	17	39.260577	-76.553062
17	2011-01-01T00:00:00	302	511	367912001	128	003669730	b	11	5	18	36.865193	-76.322923
18	2010-12-31T23:58:59	111	14	538142400	0	003669935	r	0	0	19	40.596000	-74.034650
19	2010-12-31T23:58:59	207	211	366080900	127	003669935	r	6	0	20	40.437112	-73.808987

Here is the sample data for my project,

Details for Dataset:

1- BaseDateTime: the timestamp of the vessels when the data is generated the ship movement
2- COG – Course over Ground:
3- MMSI: A unique nine-digit identification number for the vessels
4- ROT – Rate of turn: The turning point right or left from 0 to 720 degrees per minute
5- SOG – Speed over ground: Speed of the Vessels – from 0 to 102 knots and 1 knots = 1.852 km/h
6- Status: Status for each vessel
7- VoyageID: Unique Voyage ID
8- Latitude: The latitude of each Voyage
9- Longitude: the longitude of each Voyage

Analysis & Design: For my project I am analyzing few things like finding out the trajectory stops from where it started, and where it stops for some time, so I can segment the trajectory data into multiple sub trajectories. Probably my first step would be to figure out how can I use the multiprocessing for this project, so I can efficiently execute multiple tasks parallelly. Second step would be to find out the stops based on time and distance, where I can set up the distance threshold to segment it into sub trajectory. Such as where the voyage stayed and for how long it was there, or it was moving continuously with some speed.

Scalability Challenges: The data I am working on is a large dataset, and for this large data my machine is not able to handle the processing and analyzing. I need to scale the processing such as through multiprocessing where I can distribute the data on multiple cores in order to enhance the execution power.

Originally my data is in gdb format, which is around 3-GB, which my machine is unable to convert the parquet table format, so I can use the multiprocessing on the server to convert the data from gdb to parquet and can use it for my further analysis

Implementation: I am going to use python 3.6 language for my project which includes some python libraries. Pandas, numpy, multiprocessing libraries, I might be using the Spark library also PySpark

Project Plan: I will be showing the multiprocessing and converting datafiles from gdb to parquet by the progress report due date, after that I will do further processing for the project such as finding stops and segmenting trajectories into sub trajectories.