Insights from data exploration and storage and query speed explained

From my data exploration step I was able to see the type of data that is stored in each of the field and how the data is organized.

I need to fix the type of data that is stored in the 6th column the data type is mixed up.

When I ran a pandas script I found the type of data in each field.

VendorID float64

tpep\_pickup\_datetime object

tpep\_dropoff\_datetime object

passenger\_count float64

trip\_distance float64

RatecodeID float64

store\_and\_fwd\_flag object

PULocationID int64

DOLocationID int64

payment\_type float64

fare\_amount float64

extra float64

mta\_tax float64

tip\_amount float64

tolls\_amount float64

improvement\_surcharge float64

total\_amount float64

congestion\_surcharge float64

To clean up the data I think it would be best discard the vendorID, and store\_and\_fwd\_flag fields. This information will not hold much value to the analyst in solving their problems. With that being said this will also solve the data type error

That we are getting on the 6th column which corresponds to the store\_and\_fwd\_flag fields.

We will need to add a new table to show Rate codes and what the different numbers mean.

https://www1.nyc.gov/assets/tlc/downloads/pdf/data\_dictionary\_trip\_records\_yellow.pdf

This website is really important because it shows what is being stored in each of the different fields.

I think it would be best to create also create a new table with where the LocationID are referring to- what part of

New York so it makes it easier to see where the data is pointing to.

The best way to compress the files and data would be to use lossless compression method.