The block diagram below shows the workflow of my SCRUB.py file. It was found that all the timestamps were in right format and out of order; prices and units traded had negative and zero values. Some of the price and units values were outliers with very high negative/ positive values. According to this data, we do not need to check for timestamps format. Since future data can have formatting issues, this check is included.

Ticks identified as noise if:

1. Numbers of items in a tick are less than 3 or items not in right format

2. Price or units are zero or negative

3. Difference between current timestamp and base timestamp is more than three seconds. This is based on my readings about network latency that occurs in milliseconds.

My code does not identify ticks with very high price/ units as outliers since it is not the right approach. There may be genuine reason for them to be that high, so it should be investigated before categorizing them as noise.

The code is scalable for large datasets. I ran it on my laptop. For data-small.txt it takes around 3 seconds when run with logging in info mode (includes times and memory heap) or debug mode (also includes reason for a tick to be noise). However, when line-by-line memory profiling for functions is done it takes more time. For data-big.txt it takes much more time but runs to completion. Had Penzias be up I could have checked for any issues it may have if it would have taken longer to run on Penzias too. For now, I cannot be sure if it takes longer to run because of system limitations or some issue with my program.