For each file inspect the files to look for missing data, and differently formatted data.

I used excel to open each file (for a quick visual check) however python code could be written to do exploration (usually via Juypter Notebook as you can run steps at a time rather than the whole program)

For the nmi\_info.csv file

There were missing / blank lines which had to be handled reading in.

There are also entries in this file where there is no corresponding data file. Warning messages to appear. Skip processing of that datafile NMIM2.csv.

With the individual data files

There isn’t a consistent format for units’ kwh and kWh were found - lower case values.

With the different unit grade this needed to be modified to the same unit. kWh was the most common, so everything converted to that.

Missing Date Times records were found. These are just removed because there is no “guaranteed” method of extrapolating the data because we can’t guarantee the data is always given to us in sorted order.

Numeric fields look to be in range (i.e., no outliers) could have written box plots in Juypter Notebooks to check this.

Process flow

Open up nmi\_info

For each record

Read in each file

Clean in data

Standardise units.

Covert dates to local time

Consolidate into one File and load in final data store.

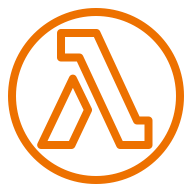
Analysis Flow

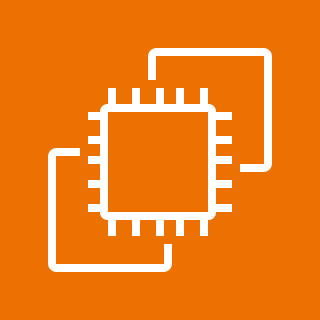
Add analysis variables Day of Week and Hour of Day

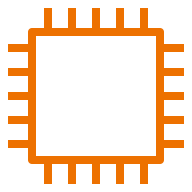
Aggregate Data by nmi, day\_of\_week, hour of day. I did a mean, but a sum would still work.



AWS Architecture







EC2 Instance (run process)

S3 Bucket with data

RedShift