

In [10]: *# The code was removed by Watson Studio for sharing.*

Extract Transform Load

This document is the ETL for the Advanced Data Science Capstone. The data is from a Kaggle dataset on water pump sensors. The goal is to read in the csv data file and transform it into a pandas dataframe.

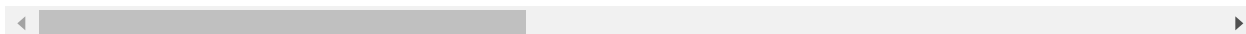
Data is extracted from the IBM Cloud Object Store

In [1]: *# The code was removed by Watson Studio for sharing.*

Out[1]:

	Unnamed: 0	timestamp	sensor_00	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06
0	0	2018-04-01 00:00:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	16.000000
1	1	2018-04-01 00:01:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	16.000000
2	2	2018-04-01 00:02:00	2.444734	47.35243	53.2118	46.397570	638.8889	73.54598	16.000000
3	3	2018-04-01 00:03:00	2.460474	47.09201	53.1684	46.397568	628.1250	76.98898	16.000000
4	4	2018-04-01 00:04:00	2.445718	47.13541	53.2118	46.397568	636.4583	76.58897	16.000000

5 rows × 10 columns



```
In [2]: #Id column is renamed for readability and the sensor_15 column is dropped as it is
data.rename(columns={data.columns[0]: 'Id'}, inplace = True)
data.drop('sensor_15', axis = 1, inplace = True)
data.head()
```

```
Out[2]:
```

	Id	timestamp	sensor_00	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06
0	0	2018-04-01 00:00:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	13.41146
1	1	2018-04-01 00:01:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	13.41146
2	2	2018-04-01 00:02:00	2.444734	47.35243	53.2118	46.397570	638.8889	73.54598	13.32465
3	3	2018-04-01 00:03:00	2.460474	47.09201	53.1684	46.397568	628.1250	76.98898	13.31742
4	4	2018-04-01 00:04:00	2.445718	47.13541	53.2118	46.397568	636.4583	76.58897	13.35359

5 rows × 54 columns

```
In [3]: pd.to_datetime(data.timestamp)
data['timestamp'] = [datetime.strptime(x, '%Y-%m-%d %H:%M:%S') for x in data.timestamp]
```

```
In [4]: data.loc[data.machine_status == 'BROKEN', 'machine_status'] = 'RECOVERING'
data.machine_status.value_counts()
```

```
Out[4]: NORMAL          205836
RECOVERING          14484
Name: machine_status, dtype: int64
```

```
In [5]: data.isna().sum()
```

```
Out[5]: Id                0
timestamp                0
sensor_00              10208
sensor_01               369
sensor_02               19
sensor_03               19
sensor_04               19
sensor_05               19
sensor_06               4798
sensor_07               5451
sensor_08               5107
sensor_09               4595
sensor_10               19
sensor_11               19
sensor_12               19
sensor_13               19
sensor_14               21
sensor_16               31
sensor_17               46
sensor_18               46
sensor_19               16
sensor_20               16
sensor_21               16
sensor_22               41
sensor_23               16
sensor_24               16
sensor_25               36
sensor_26               20
sensor_27               16
sensor_28               16
sensor_29               72
sensor_30               261
sensor_31               16
sensor_32               68
sensor_33               16
sensor_34               16
sensor_35               16
sensor_36               16
sensor_37               16
sensor_38               27
sensor_39               27
sensor_40               27
sensor_41               27
sensor_42               27
sensor_43               27
sensor_44               27
sensor_45               27
sensor_46               27
sensor_47               27
sensor_48               27
sensor_49               27
sensor_50              77017
sensor_51              15383
machine_status          0
dtype: int64
```

```
In [6]: values = {}  
        for i in data.columns[2:-1]:  
            values[i] = data[i].mean()  
  
        data.fillna(value = values, inplace = True)
```

```
In [7]: data.isna().sum()
```

```
Out[7]: Id                0
timestamp              0
sensor_00              0
sensor_01              0
sensor_02              0
sensor_03              0
sensor_04              0
sensor_05              0
sensor_06              0
sensor_07              0
sensor_08              0
sensor_09              0
sensor_10              0
sensor_11              0
sensor_12              0
sensor_13              0
sensor_14              0
sensor_16              0
sensor_17              0
sensor_18              0
sensor_19              0
sensor_20              0
sensor_21              0
sensor_22              0
sensor_23              0
sensor_24              0
sensor_25              0
sensor_26              0
sensor_27              0
sensor_28              0
sensor_29              0
sensor_30              0
sensor_31              0
sensor_32              0
sensor_33              0
sensor_34              0
sensor_35              0
sensor_36              0
sensor_37              0
sensor_38              0
sensor_39              0
sensor_40              0
sensor_41              0
sensor_42              0
sensor_43              0
sensor_44              0
sensor_45              0
sensor_46              0
sensor_47              0
sensor_48              0
sensor_49              0
sensor_50              0
sensor_51              0
machine_status        0
dtype: int64
```

```
In [11]: project.save_data(file_name = 'data_clean.csv', data = data.to_csv(index=False))
```

```
Out[11]: {'file_name': 'data_clean.csv',  
          'message': 'File saved to project storage.',  
          'bucket_name': 'fundamentalsofscalabledatascience-donotdelete-pr-9lqqxd4zzrrym  
c',  
          'asset_id': 'dcddf2ea-26da-47b3-8bc8-508f10848b17'}
```