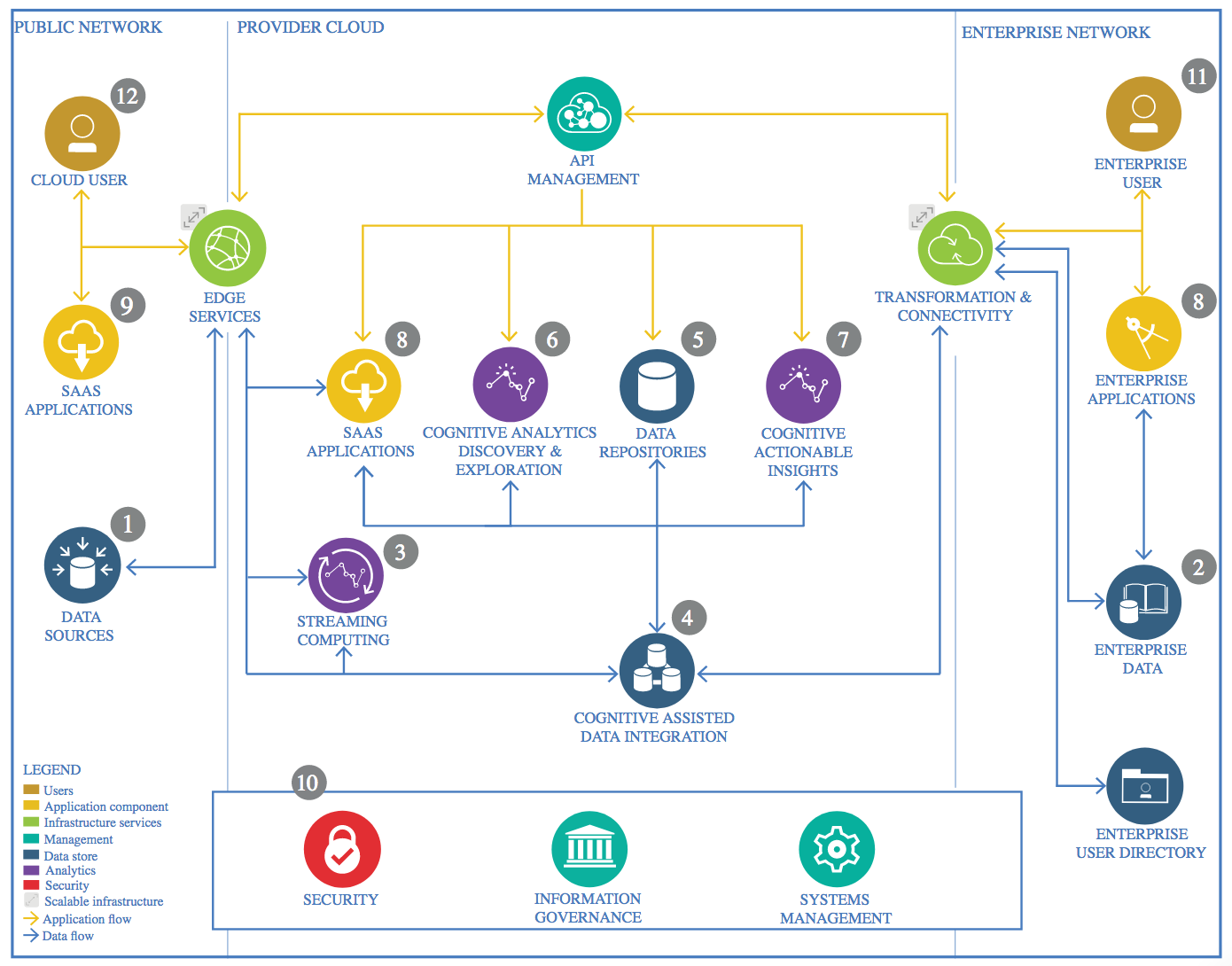
Architecture Design Document –

Soil Moisture Prediction

* Sivasakthi Jayashankar
* May 2019

# Architectural Components Overview

Overview of different stages provided by IBM is used as a guideline



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Objective –

Predict soil moisture using

## Data Source

### Technology Choice

Data was obtained from the Library of United States Department of Agriculture

* [Dataset](https://data.nal.usda.gov/dataset/data-eleven-years-mountain-weather-snow-soil-moisture-and-stream-flow-data-rain-snow-transition-zone-johnston-draw-catchment-reynolds-creek-experimental-watershed-and-critical-zone-observatory-usa-v11) includes hourly hydro-meteorological variables including soil moisture, air temperature and relative humidity from 11 sites in Reynolds Creek in southwestern Idaho

Two Datasets were retrieved.

1. Weather Dataset
2. Soil Moisture Dataset



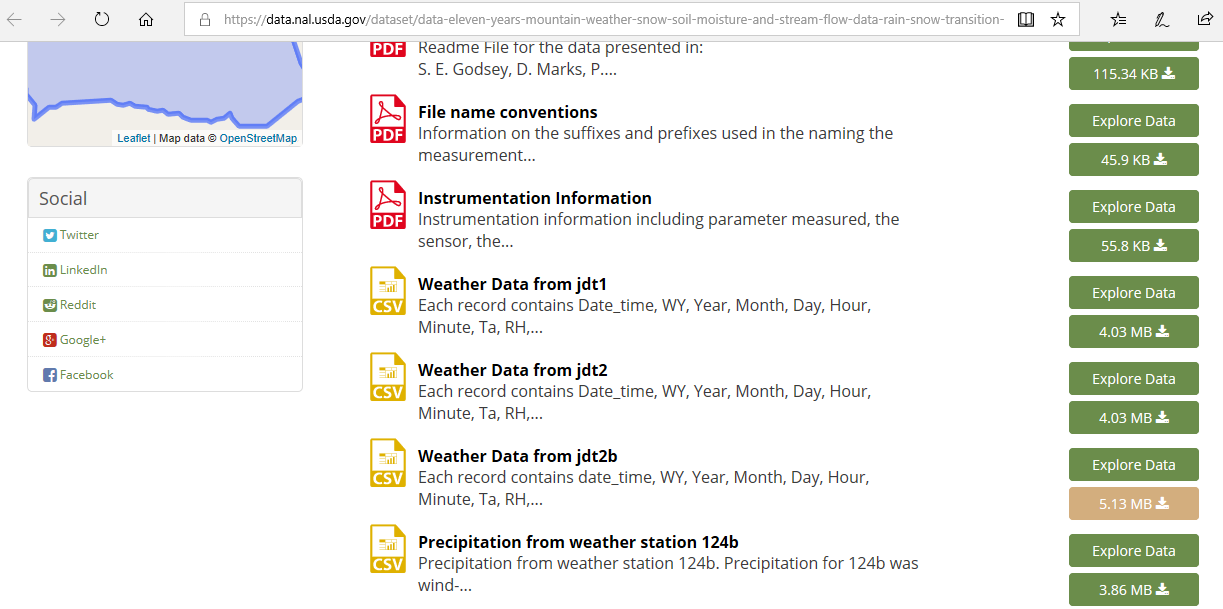
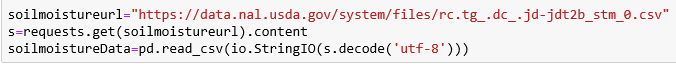


Figure 1 – Download Weather Dataset from Source



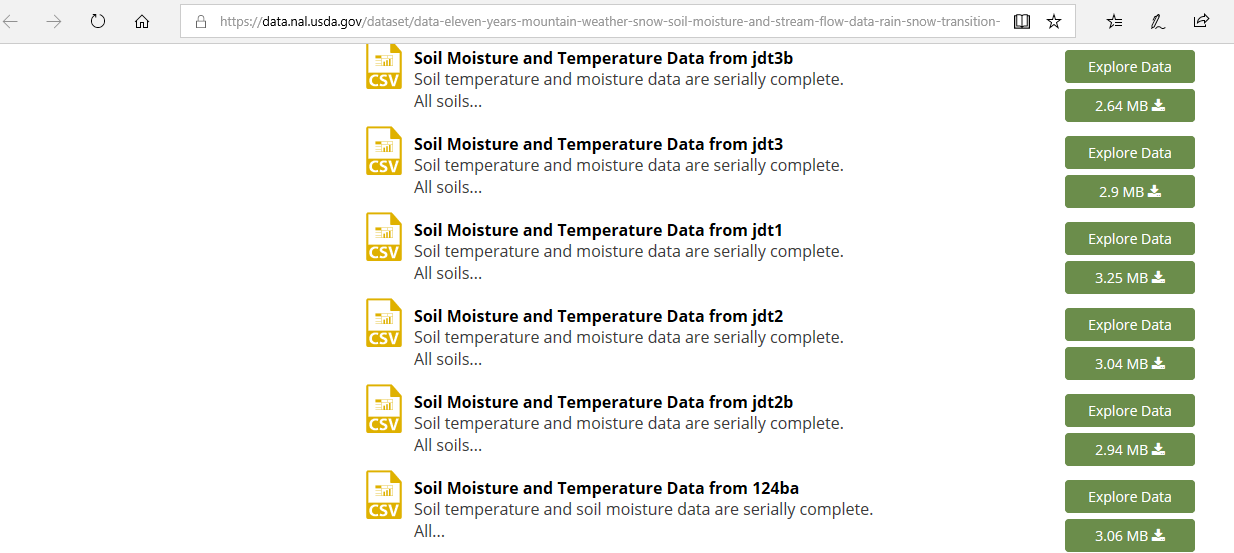


Figure 2 –Download Soil Moisture Dataset from source

### Justification

Dataset was downloaded as CSV from the USDA repository

## Enterprise Data

### Technology Choice

N/A

### Justification

Data obtained is from public data source

## Streaming analytics

### Technology Choice

Real-time data processing is not applicable for the current application

### Justification

Use case is to analyze finite data set.

## Data Integration

### Technology Choice

Apache Spark, Python, Jupyter Notebook

1. Weather Data

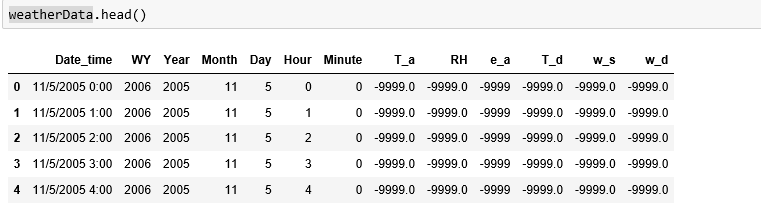


Figure 3 – Glimpse of Weather Dataset

Data Dictionary of Weather Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Feature Description | Feature Name | Feature Description |
| Date\_time | Date followed by Time | Hour | Hour of Day |
| WY | Water Year | T\_a | Air Temperature(Degrees) |
| Year | Calendar Year | RH | Relative Humidity |
| Month | Month of Year | e\_a | Water Vapor Pressure |
| Day | Day of Month | w\_s | Wind Speed |
| T\_d | Dew Point Temperature(Degrees) | w\_d | Wind Direction |

1. Soil Moisture Data

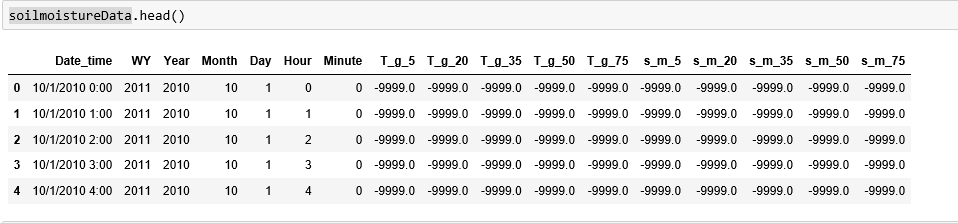


Figure 4 – Glimpse of Soil Moisture Data

Data Dictionary of Soil Moisture Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Feature Description | Feature Name | Feature Description |
| Date\_time | Date followed by Time | WY | Water Year |
| Year | Calendar Year | T\_g\_50 | Soil temperature at 50 cm Depth |
| Month | Month of Year | T\_g\_75 | Soil temperature at 75 cm Depth |
| Day | Day of Month | s\_m\_5 | Soil moisture at 5 cm Depth |
| Hour | Hour of Day | s\_m\_20 | Soil moisture at 20 cm Depth |
| T\_g\_5 | Soil temperature at 5 cm Depth | s\_m\_35 | Soil moisture at 35 cm Depth |
| T\_g\_20 | Soil temperature at 20 cm Depth | s\_m\_50 | Soil moisture at 50 cm Depth |
| T\_g\_35 | Soil temperature at 35 cm Depth | s\_m\_75 | Soil moisture at 75 cm Depth |

1. Data Cleansing

Data source documentation indicates if there was no data available, it is denoted by 9999 .

Remove all rows where variables are equal to 9999



Figure 5 – Data Cleansing – remove rows with values equal to -9999

Merge on Weather and Soil Moisture Datasets

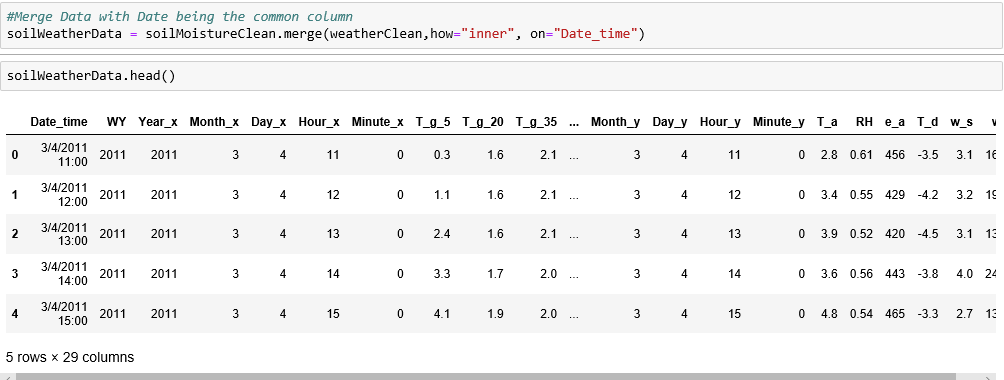


Figure 6 – Merge Weather and Soil Datasets

Drop columns not required and are duplicates from weather and soil datasets



Figure 7 – Drop Obvious Features not required for prediction

Check on values and understand the data attributes

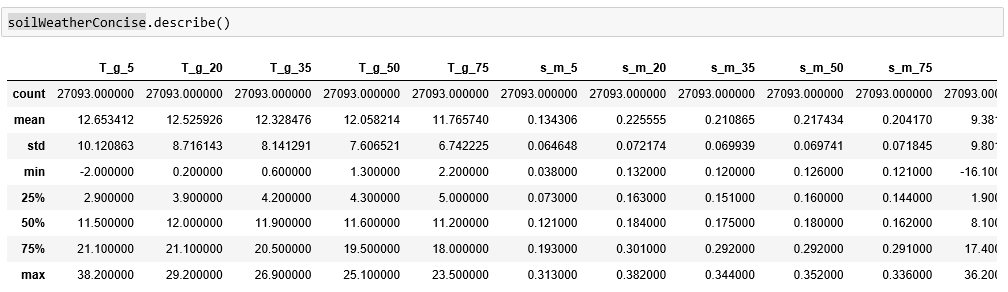


Figure 8 – Understand the values , variance and deviation

### Justification

In this stage, data is cleansed, transformed,

## Data Repository

### Technology Choice

Object storage / File system

### Justification

Object storage allows to store unlimited amount of Data. Also suited for archival when size of data becomes an issue

## Discovery and Exploration

### Technology Choice

Apache Spark ,Jupyter, Python 3.6, Matplotlib, Seaborn libraries suffice our needs for data discovery and exploration.

Visualization helps discover the relationship between the attributes.

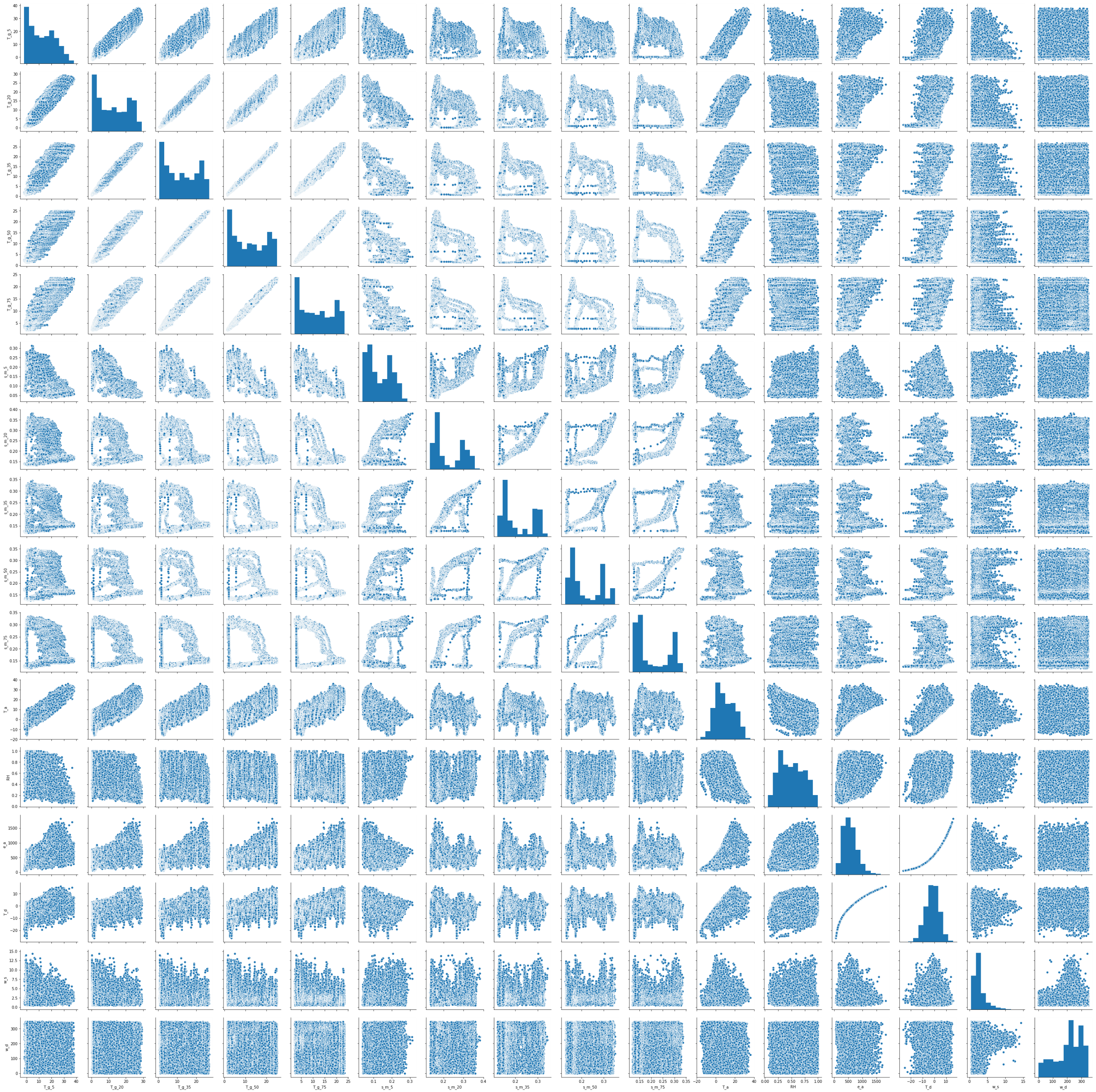


Figure 9 – Feature Relationship study

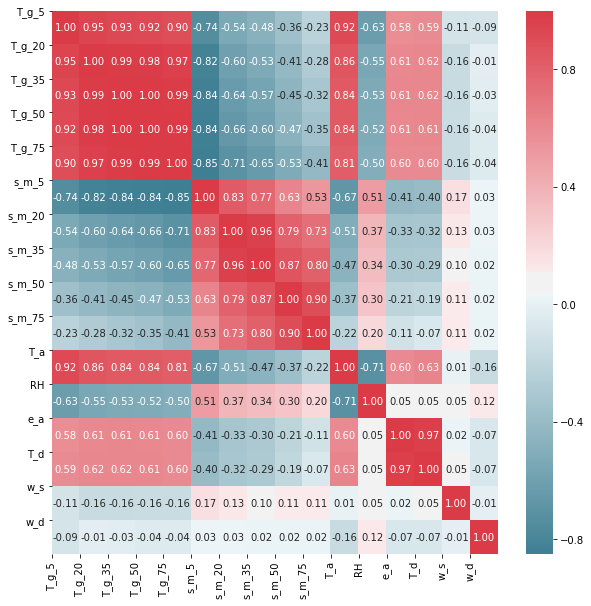


Figure 10 – Feature relation study - Correlation

### Justification

Jupyter, Python, Pandas, Seaborn & Matplotlib libraries are all open source and supported in all platforms.

## Actionable Insights

### Technology Choice

Support Vector Regression and Neural Network are the two ML models evaluated.

Support Vector Regression - Python scikit-learn library is used to build SVR model. There are different parameters in SVR. This project will compare the results between different kernels, C values, epsilon values.

Neural Networks - Sequential Keras Model with multiple hidden layers applied using Adam optimizer and relu activation.

### Justification

Exploration and Visualization of data shows data is not linear. SVR help transform to a higher dimensional data so the points can be linear separable. The options of kernels in SVR are linear, poly and RBF. Choosing the right kernel and the epsilon is very important. R square metric score can be dramatically affect with different choices of kernel and epsilon.

For sequential NN model, trying different number of layers and the number of neurons in each layer helps the model to extract and combine higher order features that are part of the data.

Metrics –

For Regression model, most appropriate to use Root Mean Square Error , Mean Absolute Error and R2

## Applications / Data Products

### Technology Choice

Guideline is to use Node-RED

### Justification

Node-RED is a great tool for process flow visualization and more consumer friendly.

## Security, Information Governance and Systems Management

### Technology Choice

IBM Identity Access Management is a known choice for security and information governance.

### Justification

Identity and Access Management (IAM) integration allows for granular access control at the bucket-level using role-based policies.