User Requirements for Data Pre-processing Pipeline

Version 1.0

PREPARED BY: ALROY CHIANG

Organization: edpr sunseap

DATE: 06-03-2023

Contents

[Revision History 2](#_Toc128985388)

[Introduction 3](#_Toc128985389)

[Objective 3](#_Toc128985390)

[Requirements 3](#_Toc128985391)

[1. Raw data input 3](#_Toc128985392)

[2. Data quality 4](#_Toc128985393)

[3. Data transformation 4](#_Toc128985394)

[4. Meta data 4](#_Toc128985395)

[5. Publish data 4](#_Toc128985396)

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason for changes** | **Version** |
| Alroy Chiang | 06-03-2023 | First version | 1.0 |
|  |  |  |  |
|  |  |  |  |

# Introduction

The idea of developing a data pipeline came about after having gone through a few developments of data driven algorithms. In those developments, we have a few encounters where functions created for a development cannot be ported to another development often due to difference in the way the data is named and delimited for instance.

It is therefore hoped that a data pre-processing pipeline will provide a more unified approach to data wrangling for data analysis, ML training and data driven algorithm development. Additionally, it is anticipated that it will improve the efficiency of development processes by ensuring data quality and integrity.

# Objective

The goal is to create a data pipeline for both Windows and Linux OS that take in user supplied raw data files, process the files and deposit them into a specified location ready for further processing. The pipeline shall take care of the data quality and integrity, standardized formats as well as parameters, perform any transformation if required, and publish the cleaned files systematically.

For example, considering a large amount of logger data being retrieved from 100 sites, with each site having 6-7 loggers. Each logger records the power in 3-Phased circuits, gathering huge amounts of raw data. Unfortunately, these data consist of irregularities and inconsistencies that makes it difficult to analyze let alone read the raw data. By creating a Data Pipeline, we can easily process large amounts of raw data in the future to provide end users with clean, readable, and usable data. Users such as the EI department can apply Data Analytics and Machine Learning to identify patterns, trends to make future decisions.

# Requirements

With this document, we aimed to record all the user/business requirements for pre-processing data. The requirements captured can be grouped into 5 categories namely:

1. Raw data input
2. Data quality
3. Data transformation
4. Meta data
5. Data publication

Detail requirements for these categories are listed below:

## Raw data input

* 1. Specify a folder to place raw data files and to initiate data pipeline.
  2. Be able to handle a single data file with data spanning across multiple months.
  3. Be able to handle multi data files with segments of data within the same month.
  4. Do not need raw data files once the raw data has been processed.
  5. Retain data file and prompt user if raw data files cannot be processed.

## Data quality

* 1. Ensure dataset does not contain duplicate columns. A duplicated column is where all its values and column names are identical.
  2. Standardize column headers.
  3. Standardize delimiters to comma.
  4. Check for missing values.
  5. Check for outliers.
  6. Check for time gaps.

## Data transformation

* 1. Transform timestamp to date-time format if it is in other formats.
  2. Change linear scale to log scale.
  3. Time differencing.

## Meta data

* 1. Location where the dataset is acquired from, in the form of address and postal code.
  2. Start time and end time.
  3. Time gap statistics.
  4. Column statistics.
  5. Column missing values count.
  6. Column outliers count.
  7. Meta data text file to have the same filename as the cleaned dataset but with a txt extension.
  8. Meta data text file to be located next to the cleaned dataset.

## Publish data

* 1. Each dataset to be exactly 1 calendar month worth of readings.
  2. To provide a method to allow users to find dataset efficiently.
  3. Datasets with huge time-gaps or datasets with outliers are to be rejected and placed into a separate folder.