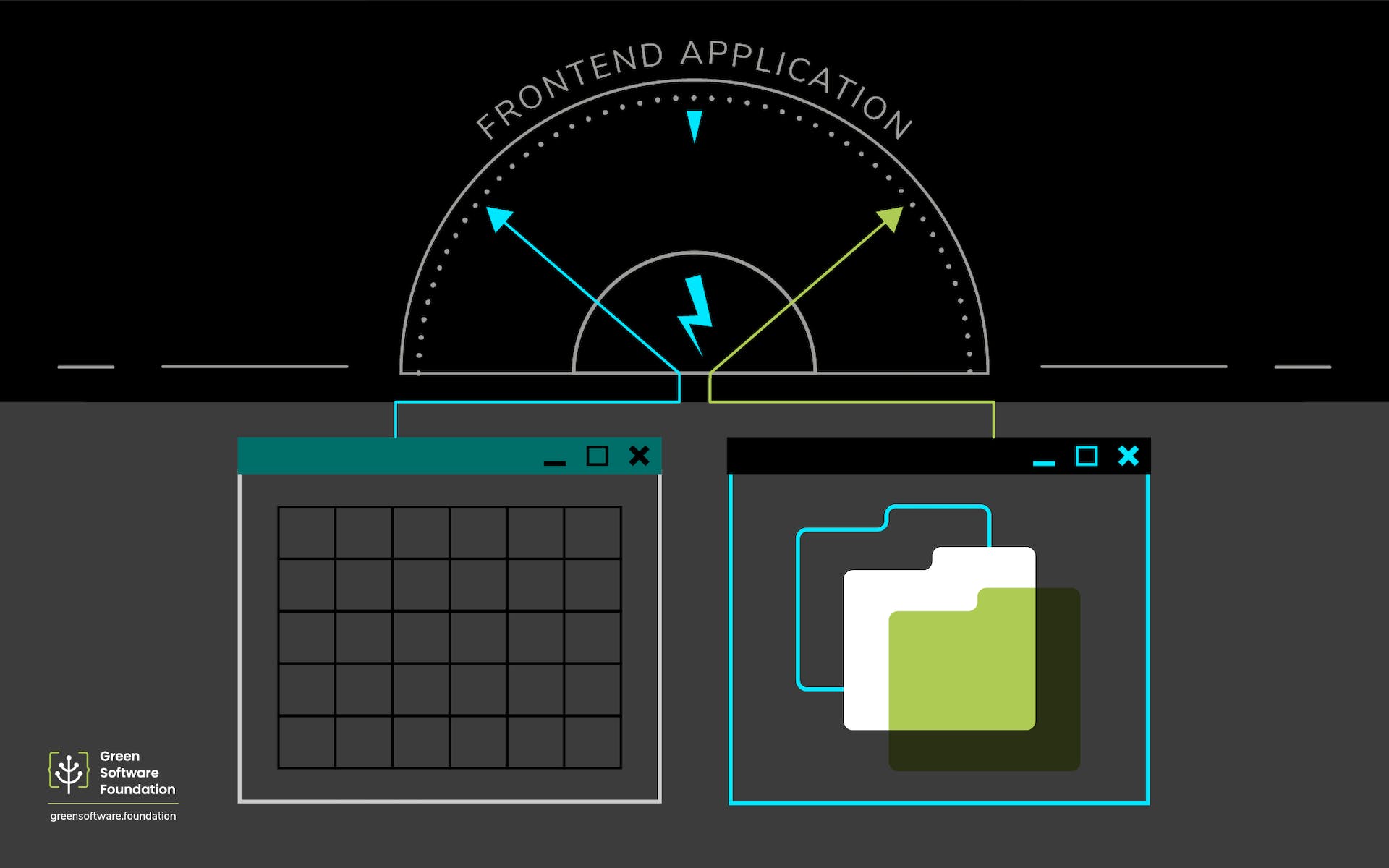
MEASURE ENERGY CONSUMPTION

* PHASE 3
* INTRODUCTION TO PREPROCESSING AND DATASETS…….



ABSTRACT

The rapid development in data science and the increasing availability of building operational data have provided great opportunities for developing data-driven solutions for intelligent building energy management. Data preprocessing serves as the foundation for valid data analyses. It is an indispensable step in building operational data analysis considering the intrinsic complexity of building operations and deficiencies in data quality. Data preprocessing refers to a set of techniques for enhancing the quality of the raw data, such as outlier removal and missing value imputation.

INTRODUCTION

This article serves as a comprehensive review of data preprocessing techniques for analyzing massive building operational data. A wide variety of data preprocessing techniques are summarized in terms of their applications in missing value imputation, outlier detection, data reduction, data scaling, data transformation, and data partitioning.

DATA PREPROCESSING

Data preprocessing is used for representing complex structures with attributes, discretization of continuous attributes, binarization of attributes, converting discrete attributes to continuous, and dealing with missing and unknown attribute values. Various visualization techniques provide valuable help in data preprocessing.

Complex structures, such as text documents, images, and graphs cannot directly be used as inputs for machine learning algorithms. They have to be described by derived attributes, such as bag of words (for text documents), various statistics (for images and graphs), or with the help of dimensionality reduction techniques, such as principal component analysis. Discretization of continuous attributes is a necessary step when using machine learning methods that cannot deal with continuous attributes. Continuous attribute values are replaced with indices of corresponding non-overlapping subintervals. Discretization algorithms need to determine both the optimal number of subintervals and the optimal boundaries between them.

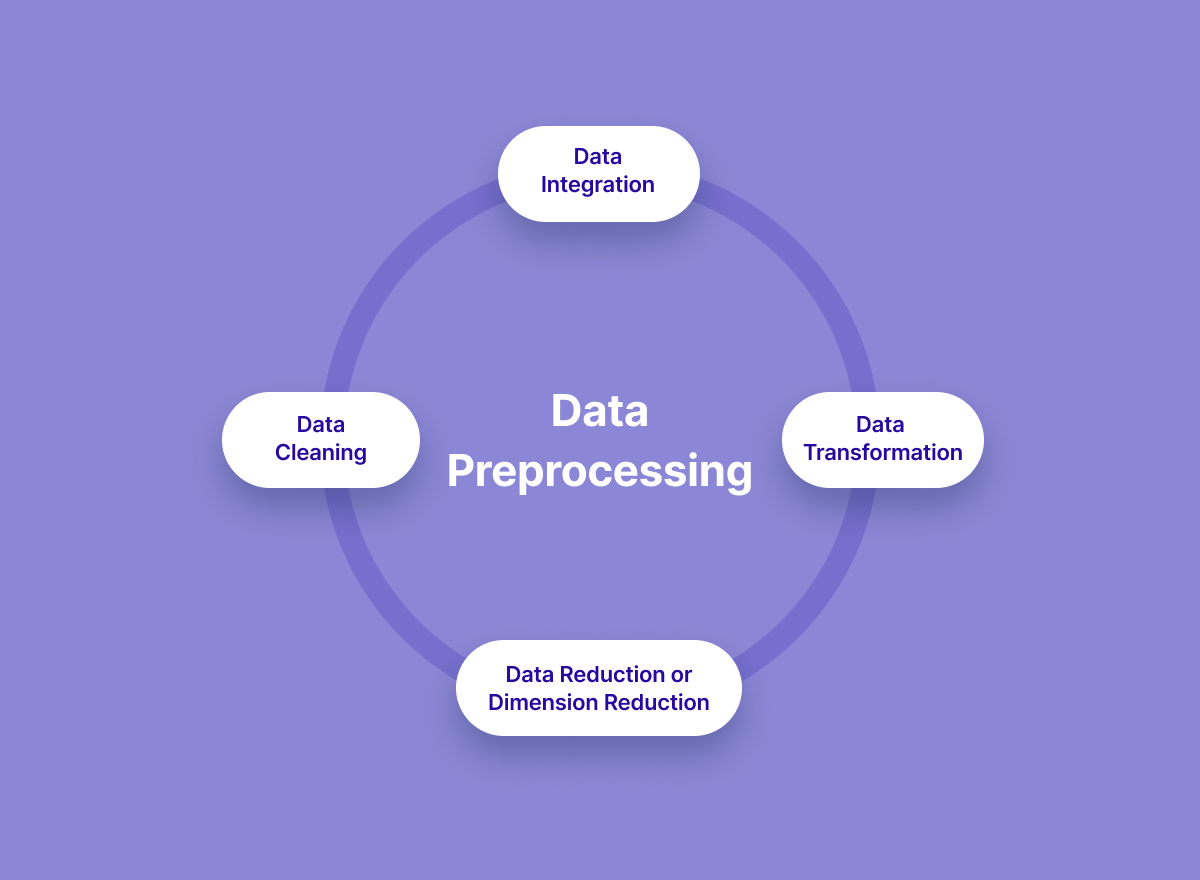
Binarization of attributes is necessary when using machine learning methods that can deal only with two-valued (binary) attributes. Continuous attributes are binarized by applying two-interval discretization. Discrete attributes are binaries either by generating a binary attribute for each original attribute value, or by generating a binary attribute for each possible different split of attribute values into two disjunction subsets. Several machine learning methods (different regression models, discriminate functions, and also neural networks) cannot deal with discrete attributes, and assume that all attributes are continuous. Multi-valued discrete attributes used within such methods are binarized, and each binary attribute is subsequently treated as a continuous attribute.

DATA PREPARATION

Preprocessing data into suitable formats is an important consideration for any analysis task, but particularly so when using MapReduce. In particular, the data must be partitioned into key/value pairs in a way that makes the resulting analysis efficient. This applies to both optionally reformatting the original data into a format that can be manipulated by R and partitioning the data in a way that supports the required analyses. In general, it is not uncommon to partition the data along multiple dimensions to support different analyses.

As a first step, it is worthwhile to convert the raw data into a format that can be quickly ingested by R. For example, converting data from a customized binary file into an R [data frame](https://www.sciencedirect.com/topics/mathematics/dataframe) dramatically reduces read times for subsequent analyses. The raw PMU data were provided in a proprietary binary format that uses files to partition the data. Each file contains approximately 9000 records, representing 5 min of data. Each record contains 555 variables representing the time and multiple measurements for each sensor.

DATA IMAGES:

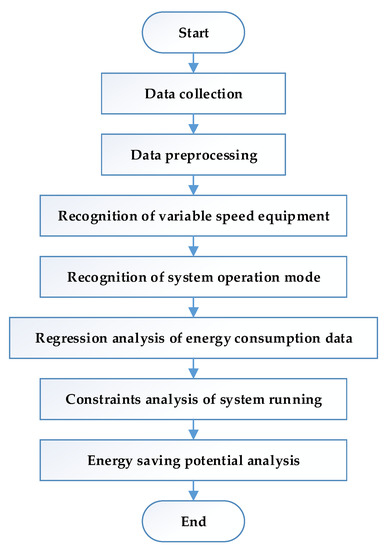


FLOWCHART

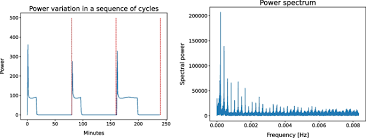
The flowchart represents the recognition of variable with the speed equipment and many tools.

Here there is a flowchart to explain us about easily and briefly.

Let us discuss it below,



GRAPH



INTRODUCTION TO DATASET

**Data Set Characteristics:**  
 Multivariate, Time-Series

**Associated Tasks:**  
 Regression, Clustering

**Data Set Information:**

This archive contains 2075259 measurements gathered between December 2006 and November 2010.  
NOTES:  
1. (global\_active\_power\*1000/60 )-

sub\_metering\_1 - sub\_metering\_2 - sub\_metering\_3) represents the active energy consumed every minute (in watt hour) in the household by electrical equipment not measured in sub-metering.

2 .The dataset contains some missing values in the measurements (nearly 1, 25% of the rows). All calendar timestamps are present in the dataset but for some timestamps, the measurement values are missing: a missing value is represented by the absence of value between two consecutive semi-colon attribute separators.

**ATTRIBUTES DATASET**  
1 . Date: Date in format

2. Time: time in format

3. Global\_active\_power: household global minute-averaged active power (in kilowatt)

4. Global\_reactive\_power: household global minute-averaged reactive power (in kilowatt)

5. Voltage: minute-averaged voltage (in volt)

6. Global\_intensity: household global minute-averaged current intensity (in ampere)

7. Sub\_metering\_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).

8. Sub\_metering\_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

9. Sub\_metering\_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

# DATA MODELS

A data model is a representation of the business data of an organization or business segment. Located on top of the physical table. You can use a model as the basis for your story. There are following types of data models:

* Acquired: Data is imported (copied) and stored in Cloud. Changes made to the data in the source system don’t affect the imported data.
* Live or Federated: Data is stored in the source system. It isn’t copied to Cloud, so any changes in the source data are available immediately if no structural changes are brought to the table or SQL view.

# DATASET

A dataset is a simple collection of data, usually presented in a table. You can use a dataset as the basis for your story, and as a data source.

Data Models complement datasets or vice versa. Datasets are more suitable for ad-hoc analysis, while data models are more suitable for governed-data use cases.

Hence therefore we had discussed about the data preprocessing and benefits of dataset.

CONCLUSION

Energy efficiency is the wave of the future. The world is quickly moving towards energy sustainability. At the same time, the mankind is trying to re-establish the connection it once had with nature. An energy efficient home is a personal step toward the direction of renewable energy, environmental protection, and sustainable living. Having such a home helps homeowners reduce their bills and provides an excellent investment. Furthermore, energy efficiency means healthier and more comfortable living that is in line with nature.

Building or upgrading to an energy efficient home requires an initial investment that is higher than the cost of a traditionally constructed home. However, there are government grants and incentives that can help to get you started and offset some of the cost. After you live in your energy efficient house for a few years, your upfront investment will pay for itself.