# INTEGRATING DATA FOR ELECTRICITY DEMAND

**GITHMIN DULSARA**
s16344

# Contents

## Introduction to Data Integration

Data integration refers to the process of combining, sharing or synchronizing data from multiple sources to provide users with a unified view.

The decision to integrate data tends to arise when the volume, complexity (big data) and need to share existing data explodes. It has become the focus of extensive theoretical work and numerous open problems remain unsolved.

Data integration encourages collaboration between internal as well as external users. The data being integrated must be received from a heterogeneous database system and transformed to a single coherent data store that provides synchronous data across a network of files for clients.

## Advantages of Data Integration

- Removes data silos.
- Enhances data consistency and completeness.
- Enables more accurate analytics and machine learning models.
- Reduces redundancy and manual reconciliation.
- Improved decision making.
- Cost savings.
- Faster reporting and analysis.

## Application Overview

An electricity generating company in Sri Lanka aims to forecast hourly electricity demand using both electricity consumption data and relevant weather data. The project involves data from 2009 to 2010, combining

- Hourly electricity demand data.
- Daily maximum/minimum temperatures and rainfall from 21 meteorological stations.
- Calendar information (month, weekday, special days).

## Variables in the Final Dataset and Justification

Temporal and Calendar based Variables:

- Date: Common identifier across all datasets
- Day_of_Week: Categorical variable to capture weekly patterns in electricity demand
- Specialty: Categorical variable distinguishing between Regular Days, Poya Days and Public/Bank/Mercantile (PBM) Holidays

Electricity Demand Variables:

- H1-H24: Hourly electricity demand values (24 columns representing each hour of the day)
- Max_Demand: Maximum electricity demand recorded during the day
- Min_Demand: Minimum electricity demand recorded during the day

These variables capture the temporal patterns of electricity consumption.

Weather Variables:

Rainfall Data:

- Katugastota_Rainfall: Daily rainfall measurements from Katugastota station
- Monaragala_Rainfall: Daily rainfall measurements from Monaragala station
- Polonnaruwa_Rainfall: Daily rainfall measurements from Polonnaruwa station
- Additional station rainfall data: Multiple other meteorological stations

Temperature Data:

- Monaragala_Max_Temp & Monaragala_Min_Temp: Daily temperature extremes
- Polonnaruwa_Max_Temp & Polonnaruwa_Min_Temp: Daily temperature extremes
- Additional station temperature data: Maximum and minimum temperatures from various stations

Weather variables are critical predictors of electricity demand.

## Data Integration Methodology

Step 1: Electricity Demand Data Processing

- Combined 48 half-hourly readings into 24 hourly values by summing consecutive pairs
- Created aggregate measures (Max_Demand, Min_Demand) for daily analysis
- Consolidated multiple files using pandas concatenation

Step 2: Weather Data Integration

Rainfall Data Integration:

- Katugastota: Transformed wide format (days as columns) to long format with Date-Rainfall pairs
- Monaragala & Polonnaruwa: Extracted data from complex Excel structures with separate 2009-2010 sections
- Multiple Stations: Pivoted station-wise data to create individual columns per station

Temperature Data Integration:

- Individual Station Files: Processed max/min temperature files separately for each location
- Multi-Station Files: Used pivot operations to create station-specific columns

Step 3: Temporal Alignment and Merging

- Common Key: Used 'Date' as the primary key for all merges
- Sequential Merging: Performed left joins to maintain electricity demand data as the base
- Date Standardization: Ensured consistent datetime formatting across all datasets

## Missing Value Imputation

Missing Value Patterns Identified: Varying degrees of missingness across different meteorological stations.

1. Backfill Method:

- Applied station-specific backfill limits based on missing data patterns
- Monaragala variables: 59-day backfill limit
- Polonnaruwa variables: 31-day backfill limit

2. Time-based Interpolation:

- Used method='time' interpolation for weather variables to account for temporal dependencies. (This method considers the actual time gaps between observations.)

3. Linear Interpolation:

- Applied to remaining numeric variables as a fallback method
- Used limit_direction='both' to handle edge cases at dataset boundaries

## Transformations and New Variable Creation

1. Temporal Aggregation:

- Half-hourly to Hourly: Combined T1+T2 → H1, T3+T4 → H2, etc.

2. Data Restructuring:

- Wide to Long Format: Transformed day-wise weather data (columns 1-31) to Date-Value pairs
- Station Pivoting: Converted station-wise data to individual columns for each location

3. Categorical Variable Creation:

- Day_of_Week: Extracted from Date using pandas datetime functionality
- Specialty Classification:
    - Identified Poya Days
    - Identified PBM Holidays (Public, Bank and Mercantile holidays)
    - Classified remaining days as Regular Days

4. Data Standardization:

- Missing Value Codes: Converted weather station codes (-9.9M, .0T) to appropriate values
- Date Standardization: Unified date formats across all data sources

5. Data Quality Improvements:

- Value Standardization: Replaced 'Tr' (trace amounts) with 0.0 for rainfall data
- Invalid Date Removal: Eliminated impossible dates (e.g., February 30th) during date creation