**Problem Statement**

To build a regression methodology to determine Predict the power that is generated (in KW/h) based on the various features provided in the dataset.

**Architecture**



**Data Description**

The data contains the following attributes:

**Features:**

1. **tracking\_id:** Represents a unique identification number of a windmill
2. **datetime:** Represents the date and time of a record.
3. **wind\_speed(m/s):** Represents the speed of wind (in meter per second)
4. **atmospheric\_temperature(°C) :** Represents the temperature (in degree Celcius) of a town or village that the windmill is present in
5. **shaft\_temperature(°C):** Represents the temperature of the shaft (in degree Celcius)
6. **blades\_angle(°):** Represents the angle of the blades of a wind turbine (in degrees)
7. **gearbox\_temperature(°C):** Represents the temperature of a gearbox (in degree Celcius)
8. **engine\_temperature(°C) :** Represents the temperature of an engine (in degree Celcius)
9. **motor\_torque(N-m):** Represents the torque of a motor (in Newton meter)
10. **generator\_temperature(°C) :** Represents the temperature of a generator (in degree Celcius)
11. **atmospheric\_pressure(Pascal) :** Represents the atmospheric pressure (in Pascals) in that area
12. **area\_temperature(°C) :** Represents the temperature (in degree Celcius) of the area within a 100 m radius of the windmill
13. **windmill\_body\_temperature(°C) :** Represents the temperature of the body of a windmill (in degree Celcius)
14. **wind\_direction(°) :** Represents the direction of the wind (in degrees)
15. **resistance(ohm) :** Represents the resistance against the wind
16. **rotor\_torque(N-m) :** Represents the torque of a rotor (in Newton meter)
17. **turbine\_status :** Represents the status of the turbine (masked)
18. **cloud\_level :** cloud in the sky on a particular day Extremely low, Low, Medium
19. **blade\_length(m) :** Represents the length of the blades of a windmill (in meter)
20. **blade\_breadth(m) :** Represents the breadth of the blades of a windmill (in meter
21. **windmill\_height(m) :** Represents the height of the blades of a windmill (in meter)

**Target Label:**

1. **windmill\_generated\_power(kW/h) :** Represents the power generated (in Kilo Watt per hour)

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1. Name Validation- We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."
2. Number of Columns - We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad\_Data\_Folder."
3. Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".
4. The datatype of columns - The datatype of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad\_Data\_Folder".
5. Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad\_Data\_Folder".

**Data Insertion in Database**

1. Database Creation and connection - Create a database with the given name passed. If the database has already been created, open a connection to the database.
2. Table creation in the database - Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file. If the table is already present, then the new table is not created, and new files are inserted in the already present table as we want training to be done on new as well as old training files.
3. Insertion of files in the table - All the files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".