Performance Analysis & Algorithm Optimization

- **Programming Languages: -** Python 3.0 & R.
- **IDE (Integrated Development Environment): -** Zeppelin.
- **Lines of Code: -** 110.
- **Packages Used: -** Numpy, Pandas, Matplotlib, Scipy, org, java.

Operation	**Bottleneck**	**Solution**	**Time Taken**
Package Loading	Loading complete module is time consuming. And in this application packages used comprises of many functions and modules. This makes it bulky.	Only load required libraries or functions from the module. Like: **import matplotlib.pyplot.** Instead of loading complete matplotlib, only accessed pyplot.	45 seconds
Data Loading	Data is in .csv format and has more than 5 columns. Want to read data properly and with headers.	**Pandas** make our life easy. We can read data with simple and optimized function **read_csv().**	1 second
Data Concatenation	As source data is divided by month. For performing data analysis or visualization need to combine the complete data into one set. Need a solution to merge all the divided datasets in one operation.	**Pandas.concat()** function can perform data merging row as well as column wise.	1 second

Data Cleaning	Need to remove skewness and fill null values in big data set.	Created a function to fill the null values with mean and remove skewness from the data. Functions help in code reusability.	1 second
Data Exploration	1. Need insight of data like: mean of all the columns, maximum value of columns, length of dataset, standard deviation etc. Cannot perform each operation individually as it will be time consuming. 2. Want to understand correlation of parameters used in the study.	1. **Dataset.describe()** Function provides complete story of the dataset. Instead of performing individual operation, we can get complete story using this one function. 2. **Pearsonr** function from **scipy** library helps to produce correlation.	1. 3 seconds. 2. 1 second.
Data Visualization	Want graphical exploration of data. It would be time consuming to design model with all the columns in the dataset.	**Matplotlib.pyplot** Provides easy to understand graphical view of dataset.	2 seconds