**Performance Analysis**

**Number lines of Code**: 102

**Languages**: Py Spark and R.

**Software**: Zeppelin

**Code Performance:**

The data had 419 files in total, each file representing the longitude and latitude of a certain location. It took long time to perform operations on the data due to the size of the data set, so we aggregated the data based on the longitude and longitude. The arithmetic mean was taken of each pollutant in each file, which represented the pollutant level in that particular latitude and longitude.

**Tables Describes the Performance of the Code**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Duration** | **Language** | **Bottleneck** | **Comments** |
| Loading of Data | 1 min | PySpark | No Bottlenecks |  |
| Running clustering | 5-6 mins | PySpark /R | Was Too long to run the cluster and the cluster did not make meaningful result | Aggregated the data based on longitude and latitude |
| Aggregate Data | 5 mins | PySpark | No bottle neck |  |
| Running clustering on aggregated data | 2 mins | R | No Bottle neck | Aggregating the data set helped in reducing the clustering time |
| Plotting the graphs | 1 min | R | No Bottle neck |  |