Joanna Riascos

Dr. Jaume

Performance Analysis

March 26th, 2017

Number lines of code: 195

Languages: R and PySpark

Software: Zeppelin

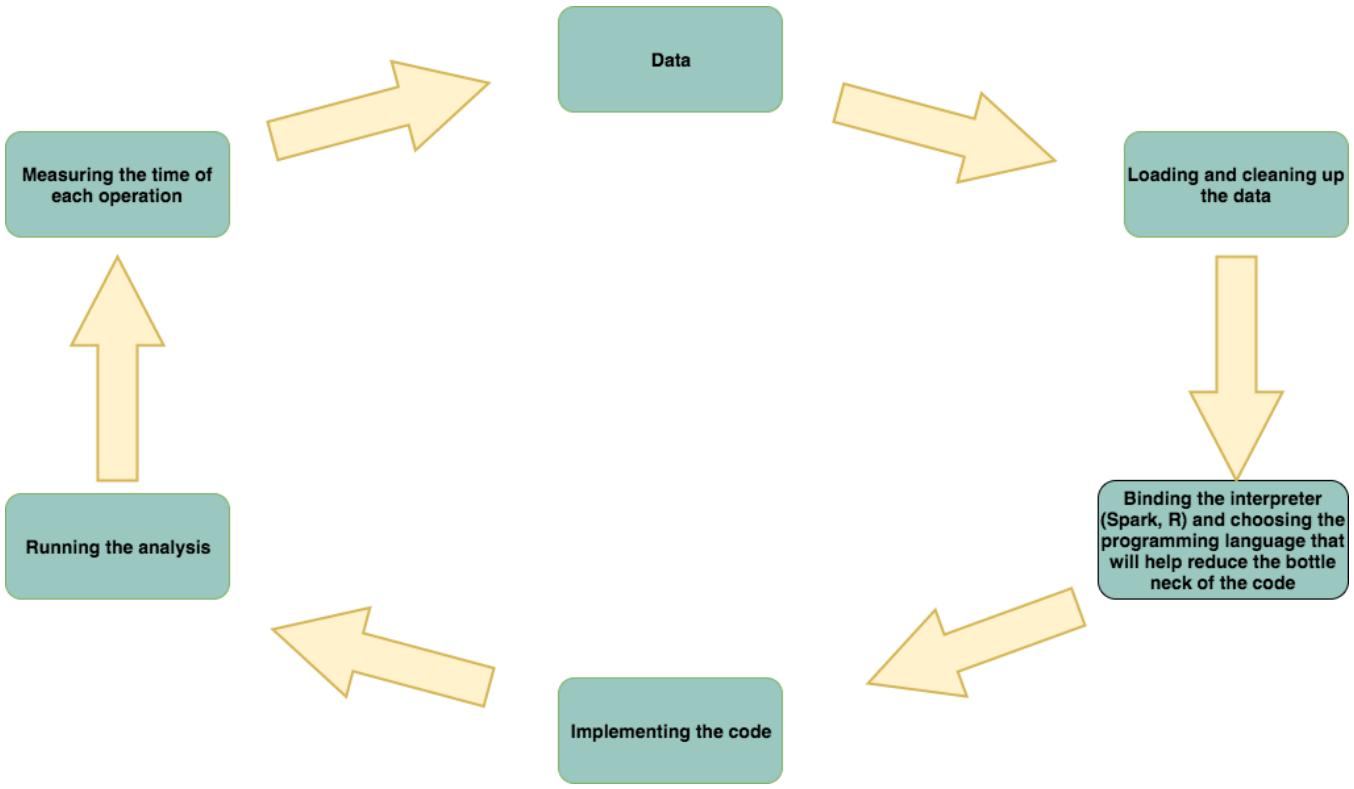
Performance Analysis

The bottlenecks that we found while running the code were the following:

- **Trying to merge the parking and pollution dataset**
- **Running the linear regression by using all of the variables**

The way that we managed to make the implementation much faster was by only choosing the most important variables and we broke down the linear regression steps.

Workflow



Description	Time Of Operation	Programming Language	Bottleneck Identified? Yes/No	Comments
Data Loading	3 mins	**PySpark**	No	
Clean-up	2 mins	**PySpark**	No	
Binding the interpreter	5 mins	**PySpark**	Yes, it took quite some time to bind the data into the system.	We changed to the R interpreter to help reduce the time and optimize our code.
Creation of RDD pairs	2 mins	**PySpark**	No	
Running the linear Regression	5 mins	**R**	Yes, the linear regression took long to run.	We chose the most important variables and split the variables.
Plots	3 mins	**R**	No	

Results

As stated before, by switching to the R interpreter we were able to make the implementation much faster. Splitting the variables and running the analysis optimized the linear regression. In the future, we would like to perform cross validation to get and analyze the accuracy of our model.