READ ME

SUBMITTED CODE

- **Project Code:** This code is different from out previous code. Using spark configurations, broadcasting, caching and improvements in algorithms this new code is optimized.
- **Filtering Code:** We prepared a code to extract data from files given at http://spatialhadoop.cs.umn.edu/datasets.html**.**

MONITORING

We used following two tools to measure performance: -

- Dstat on linux.
- Activity monitory on mac osx.

'dstat' is a tool used to monitor jobs on OS level. We run the following command to get required monitoring data and export it into csv:

```
dstat -tcmn -N wlan0 --output filename.csv
```

where filename.csv is the name of the desired output .csv file
Iif you are connected via LAN the following command should be ran:

```
dstat -tcmn -N eth0 --output filename.csv
```

REPORTING

Filename: Report.xlsx

- Sheet 1: contains measurements for single node, two node and four node clusters.
- Sheet 2: contains node configurations.