Assignment 06

Neha Patwardhan, Joy Machado

February 25, 2016

Assignment 06 Report

This is the markdown created to compare the time taken to calculated the missed connections between flights on hadoop distributed system and spark distributed system.

The programs were run in psuedo mode and on the EMR to find the time taken to calculate the number of missed connections

The R script is written to compare the taken to run all the programs on various modes

This is the time taken to run file 55.csv on all the modes

```
require(ggplot2)
```

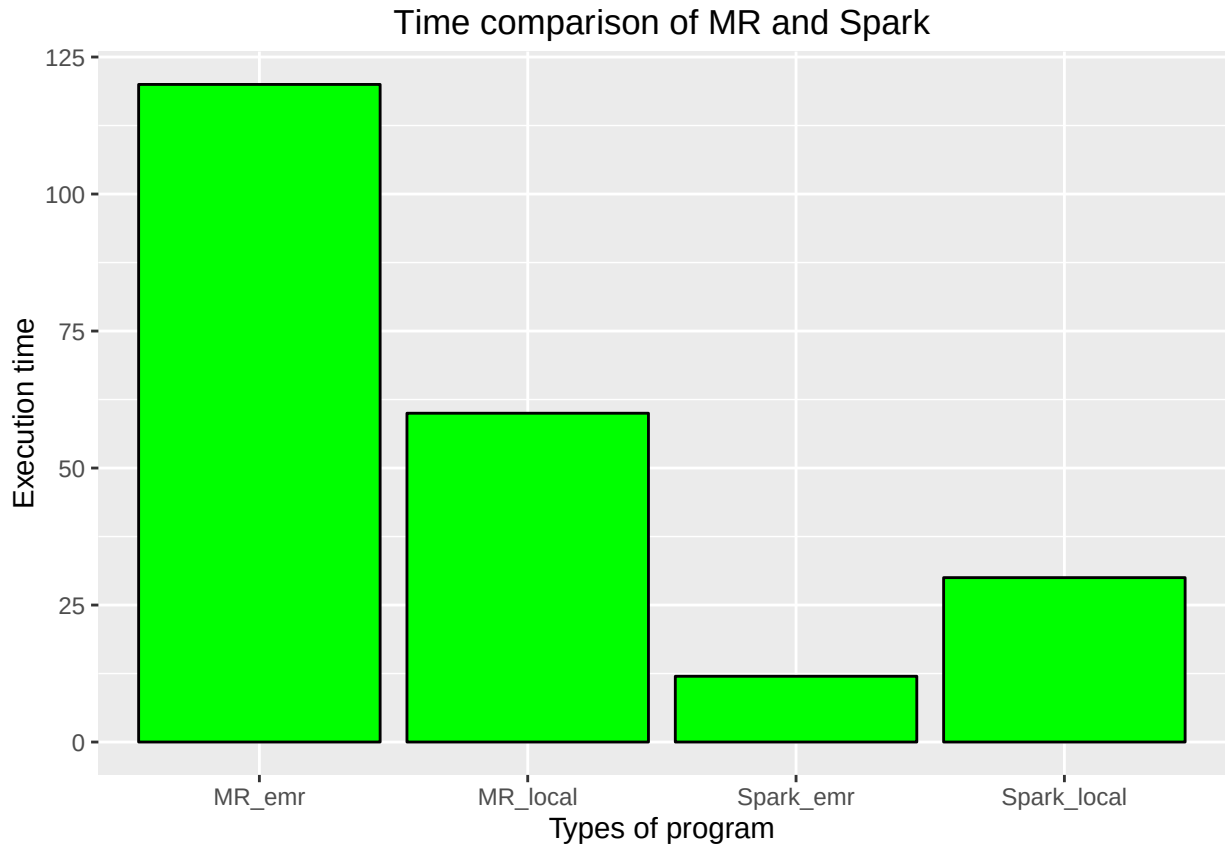
```
## Loading required package: ggplot2
```

```
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
args <- commandArgs(TRUE)
srcFile <- "C://Users//Neha//Desktop//Assignment06//output.csv"
data <- read.csv(srcFile,header =FALSE,sep="\t")
names(data) <- c("TypeOfProgram","Time")
```

The Plot to compare the time taken by the program to run



Conclusion

As you can see, it took us a large amount of time to run the program on EMR and psuedo distributed

However Spark reduces the over head by a very very large amount and is a much better distributed system to run large amounts of data