**Spark RDD API Assignments**

Input Data: log\_file.txt

**Sample Data**

"date","time","size","r\_version","r\_arch","r\_os","package","version","country","ip\_id"

"2012-10-01","00:30:13",35165,"2.15.1","i686","linux-gnu","quadprog","1.5-4","AU",1

"2012-10-01","00:30:15",212967,"2.15.1","i686","linux-gnu","lavaan","0.5-9","AU",1

"2012-10-01","02:30:16",167199,"2.15.1","x86\_64","linux-gnu","formatR","0.6","US",2

"2012-10-01","04:06:10",1023,NA,NA,NA,NA,NA,"US",4

Assignment1:

1. Use the log\_file as input
2. Filter out records where package = “NA” and version = “NA”
3. Find total number of downloads for each package
4. Find Average download size for each Country

Assignment2:

1. Use the log\_file as input
2. Create a Scala hashmap with country codes and country names for all country codes available in the log\_file
3. Perform a broadcast join to generate the below output using RDD API

Country Code, Country Name, Total Downloads

1. Save the output as a text file

**Spark SQL Assignments**

Assignment3:

1. Read the log\_file as a RDD
2. Convert the input RDD to a Data Frame
3. Add a new column to Data Frame called Download\_Type

If Size < 100,000, Download\_Type = “Small”

If Size > 100,000 and Size < 1,000,000, Download\_Type = “Medium”

Else Download\_Type = “Large”

1. Find the total number of Downloads by Country and Download Type
2. Save output as a Parquet File

Assignment4:

1. Read the log\_file and convert to a Data Frame and save as JSON File
2. Read the JSON file into a data frame
3. Perform the following transformations using Data Frame API
   1. Group By Date
   2. Find Max, Min and Average Download Size for each Date
   3. Sort records by Date in Ascending Order
   4. Save the output as a JSON file

**Spark Streaming Assignment**

Assignment5:

Create and execute a Spark Streaming program which will do the following

1. Read log records using a Socket Text Stream with a Batch Interval of 10 seconds
2. For each RDD Batch, Filter out records where package = “NA” and version = “NA”
3. Save records with size > 1,000,000 logs\_large\_downloads folder and all other records to logs\_others folder