**Click Stream Data Analysis**

**Purpose:**

Objective of this project is to collect click stream data of USA Government websites which is high in volume and velocity, and store it to analyse in a cost effective manner for enhanced insight and decision making.

**Method:**

**Step 1:**

In the below link, click stream data for a period of 1 year is archived in to numerous small files and provided.

http://www.usa.gov/About/developer-resources/1usagov.shtml

**Step 2:**

As there are numerous files, we unzipped all the files on Linux platform and then concatenated in to a single file.

**Step 3:**

Now we loaded this data on to HDFS.

**Step 4:**

Data provided is in JSON Format. And In this project our interest is find top 10 websites per country and per month.

So the fields we require would be only url, month and country. So pre-processing this data and extract the required fields is pretty much important as this would decrease the processing time and storage on hadoop. Since this data is already loaded on to hadoop, now we can use the map reduce model to get the required fields. So it is enough to have identity reducer to store the output on to the HDFS. For this operation we could have used Pig inbuilt JsonReader function to extract the required fields, but this function is provided from pig 1.0 version, but we have a older version. Another option we have is to write a UDF and register it, but in general on hadoop python should be installed to register UDF, since we do not have python and installing that would might un stabilize the existing environment. Finally we are left with hadoop streaming. In this case we can use any language which can read from standard input and write to standard output. But we chose python, as by default it is present in Linux environment and python is very good for data processing.

In mapper phase, python will read each line from standard input, decode the JSON format and store it in a dictionary which is very similar to map (key-value) in java. Now required fields are URL, country and month, getting URL and country is simple. But getting month from time stamp is a bit complex, in which we need to decode it to human readable format and then store month in separate variable. Now emit these values from mapper. After this they undergo shuffle and sort phase, followed by reducer which writes the output to HDFS, here we use identity reducer since there aggregation is required. This step is the most important step which discards all the unnecessary data and only picks the required fields.

**Step 5:**

Data now has only three tab separated fields. Now we need to process this data and find:

1. The top 10 most popular sites in terms of clicks
2. The top-10 most popular sites for each country, and
3. The top-10 most popular sites for each month.

**Analysis:**

Results will give top URL’s visited in each country and per month.