Lab assignment 1

Group B

→ Title : Perform the operations using python on the facebook metrics dataset.

→ Objectives :

a) To understand and apply the analytical concept of big data using python.

b) To study detailed concepts of python.

→ Software requirements :

a) Ubuntu 14.04 / 14.10

b) GNC C compiler

c) Hadoop

d) Java

e) Python platform.

→ Problem statement :

Perform the following operations using python on facebook metrics dataset.

a) Create data subsets

b) Merge data

c) Sort data

d) Transposing data

e) Shape and reshape data

→ Theory :

a) Data reshaping :

- It is about changing the way data is organised into rows and columns.
- Most of the time data processing in python is done by taking the input.

- It is easy to extract data from rows and columns of a data frame but there are situations when we need the data frame in a format that is different from format in which we received it.

b) Joining rows and columns in data frame :

- We can join multiple vectors to create a data frame using the cbind() function.
- Also we can merge two data frames using bind() function.

Merging data frames : We can merge two data frames by using the merge() function. The data frames must have some column names on which the merging happens.

c) Subset :

- Subsetting vectors, matrices and data frames.
- Return subsets of vectors, matrices or data frames which meet conditions.
- This is a generic function with methods supplied for matrices data frames and vectors (including lists).
- Packages and users can add further methods.
- For ordinary vectors, the result is simply x[] subset of lisma (subset).

d) Melting and casting :

- One of the most interesting aspects of R programming is about changing the shape of data in multiple sets steps to get a desired shape.
- The functions used to do this are called melt() and cast().
- The melt function takes data in wide format and stacks a set of columns into a single column of data.
- To make use of the function we need to specify a data frame, the id variables and measured variables to be stacked. The default assumption is that all columns are not specified as id variables.

d) Sorting the data :

- To sort a data frame in R, we use the order function.
- By default, sorting is ascending prepend the sorting variable by a minus sign to indicate descending order.

→ Conclusion : Thus, we have learnt to perform different reshape operations using python.