Benchmarking Efficient Data Exchange Between Office Open XML Spreadsheet File Format and R

A. P. Awasthi

Department of Statistics, Amity University, Noida, Uttar Pradesh, India

ABSTRACT

Office Open XML Spreadsheet File also known as Excel files (*.xlsx) are most common medium for storing, exchanging, and distributing data. Post release of Microsoft Excel 2007, capability of storing data into excel file has been increased a lot and modern era Excel files can store up to 1,048,576 rows by 16,384 columns in a single worksheet. This make excel a suitable file format for data management where advance and technologically complex systems like databases are not available. Excel is a go to file format for researchers, academicians and in industry. Efficient data exchange strategies will help the end user (Researcher/data scientist/Analyst etc.) to manage the work hours effectively. This will directly translate into working efficiency of resource. Also, efficient data exchange strategies help organization in selecting the right method for data exchange activities and setting up the protocols and the guidelines. There was no prior work done to benchmark the data exchange between the Office Open Xml Spreadsheet Files and R in the past. This will be an approach to setup the baseline which can be further generalize for other format/packages/volume in the future. In R environment, there are numerous packages which deals with the data exchange between R and Excel (*.xlsx). Data exchange includes two activities, import data (from excel files to R environment) and export data (from R environment to excel files). A benchmark study to measure the efficiency of these activities in the scope of *readxl and openxlsx* for data import and *writexl and openxlsx* for data export. Dataset size of 10^2 to 10^5 (10^2, 10^3, 10^4, and 10^5 records) created from New York flight data (R package nycflights13) using simple random sampling with

DOI: 10.21467/abstracts.138 ISBN: 978-81-957605-2-7 (eBook)

replacement. Once the datasets are created, both activities (Import/Export) have been performed on each dataset in a well manged Linux cloud VPS 50 times and execution time recorded. Execution data analysed for benchmarking the efficiency of data exchange process. the hypothesis was first established using exploratory data analysis of the data followed by inferential statistics method. Benchmark for import process shows 32% efficiency of openxlsx over readxl in case of tiny dataset, while readxl found 142% (average time to process 1000 transactions) over openxlsx for large datasets. Benchmark for export process shows writexl perform 288% (average time to process 1000 transactions) efficient export capabilities over openxlsx export methods.

Keywords: openxlsx; readxl; writexl; Data Exchange in R

ISBN: 978-81-957605-2-7 (eBook)

Series: AIJR Abstracts

DOI: 10.21467/abstracts.138