Homework-2

**Out date:** June 18, 2021

**Due date:** June 27, 2021 at 11:59PM

Team#: \_\_\_

Team Member-1:Charles\_Colgan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

Team Member-2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

Team Member-3:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

**Submission**

1. Answer the following questions.
2. Rename this word file to “HW2-YourTeam#” (e.g., HW2-Team1.docx).
3. Complete Table-1.
4. Visualize performance data.
5. Upload your R script and the word document on blackboard.

The dataset contains users’ events from eCommerce website. The data has 8765005 rows and 9 columns. The size is about 1.36GB. For details, please read ‘File structure.txt’. The dataset has been modified and reduce from original data. There is no missing value. Download the compressed data (0.5 GB) from here:

<https://drive.google.com/file/d/1OtYZA_UFzgiqprExExXtvqhaNWT65jOg/view?usp=sharing>

Use R packages for big data analytics to perform the following task. In particular, provide four solutions – one R script per solution as follows:

1. **Solution-1** 🡪 Use only ff/ffbase package (for data loading and other computations)
2. **Solution-2** 🡪 Use bigmemory package (for data loading and other computations)
3. **Solution-3** 🡪 Use parallel package (for multi-core processing)
4. **Solution-4 🡪** Use data.table package (for data loading and other computations)

**Tasks**

1. Load the data. **[20 points – 5 points per solution]**
2. Calculate the average daily visitors and find which day of the week has the highest number of daily visits. **[20 points – 5 points per solution]**
3. Calculate the average price for each brand. **[20 points – 5 points per solution]**
4. Calculate the conversion rates for each brand. The conversion rate is the number of conversions divided by the total number of visitors. For example, if a product has 200 views in a month and has 50 purchase, the conversion rate would be 200 divided by 50, or 25%. **[20 points – 5 points per solution]**.

Run tasks 1-4 solution on your completer and on the AWS. Compare the computing performance of cloud computing with laptop computing for various activities listed in Table-1. Report the performance in Table-1 and visualize the performance data (via a bar chart). **[20 points – 4 points per step in Table-1]**

**Step-1 of 2**

**Table-1: Processing Speed (in sec)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Configuration** | **Laptop** | **EC2 c5.large** | **EC2 c5.xlarge** |
| CPU | 12 compute cores | 2 VCPUS | 4 VCPUS |
| RAM | 32BG | 4GB | 8GB |
| **Task-1: Reading the data** | | | |
| **Baseline performance Traditional R** \* read.table() | 32.77sec | 45.986sec | 29.703sec |
| **ff package**  read.table.ffdf() | 687.9sec | 342.825sec | 256.015sec |
| **Bigmemory package** read.big.matrix() | 246.24sec | 234.122sec | 82.043sec |
| **data.table package** fread() |  |  |  |
| **Task-2: Data structuring \*\*** | | | |
| Traditional R | 72.34sec | 68.766 | 42.251sec |
| ff package | 94.25sec | 58.324sec | 36.651sec |
| Bigmemory package | 395.42sec |  | 179.763sec |
| parallel package |  |  |  |
| data.table package |  |  |  |
| **Task-3: Average daily visitors** | | | |
| Traditional R | 0.56sec | 1.21sec | 0.766sec |
| ff package |  |  |  |
| Bigmemory package | 8.51sec |  | 6.698sec |
| parallel package |  |  |  |
| data.table package |  |  |  |
| **Task-4: Average price** | | | |
| Traditional R | 6.81sec | 8.591sec | 5.484sec |
| ff package | 36.77sec | 35.618sec | 24.486sec |
| Bigmemory package | 5.75sec |  | 4.165sec |
| parallel package |  |  |  |
| data.table package |  |  |  |
| **Task-5: Conversion rates** | | | |
| Traditional R | 5.49sec | 5.063sec | 3.396sec |
| ff package | 8.89sec | 5.769sec | 4.179sec |
| Bigmemory package | 3.4sec |  | 2.595sec |
| parallel package |  |  |  |
| data.table package |  |  |  |

**Notes:**

\* Use *Traditional R* as a baseline for the performance comparison. If traditional R is unable to handle the data (e.g., program is crashing, or taking very long time to load or process the data), just mention it in the table with a brief justification.

\*\* The data structuring involves any rearranging of the data necessary before the data analytics. This includes data merging, feature extraction (such as an extracting specific columns), etc.

**Step-2 of 2**

Provide visualizes to compare data in Table-1.

**Programming Style:** Pick one programming style and use it throughout your program. You may use Google’s R Style guide as reference:

<https://google.github.io/styleguide/Rguide.xml>

**Programming Comments:** Please make sure the code is well-commented.

Submit the R scripts to blackboard. The scripts should be self-contained, meaning each script should read the data file from the folder where the script is and produce the final outputs.