Big Data Management
Assignment 2

Description

In this assignment your task is to prepare the batch layer (off-line processing pipeline) of the lambda architecture that will enable us to perform some analytics on a dataset and then store them on Apache Cassandra for the serving layer. You are required to use Apache Spark's SQL API to compute some simple analytics. You will be using the CRAN package download logs (`http://cran-logs.rstudio.com`). These log files contain all hits to `http://cran.rstudio.com` mirror related to downloads of the R packages. The raw log files have been parsed into CSV and anonymised. Since these logs contain massive amount of data (from 2012 to date), we will only be using a recent one which represent the logs for 31st of October, 2021. This log file are available at:

`http://cran-logs.rstudio.com/2021/2021-10-31`

The package download logs contain data about the following variables:

```
date: Download date
time: Download time (in UTC)
size: Package size (in bytes)
r_version: Version of R used to download package
r_arch: Processor architecture (i386 = 32 bit, x86_64 = 64 bit)
r_os: Operating System (darwin9.8.0 = mac, mingw32 = windows)
package: Name of the package downloaded
country: Two letter ISO country code
ip_id: A daily unique id assigned to each IP address
```

Setting Up

Follow the Cassandra setup instructions provided on Moodle. You can also follow:

Official guidelines on setting up Apache Cassandra on your machine are available at:
`http://cassandra.apache.org/download/`
To configure Apache Cassandra to work with Apache Spark, see:
`https://github.com/datastax/spark-cassandra-connector`

Questions

You answer similar questions as in assignment 1 but this time you are required to use the SQL API of Apache Spark. Once the results are computed, prepare Cassandra structures/tables and Spark code that saves the results (batch views) data into these structures:

1. Show total number of downloads for packages `ggplot2 and dplyr`
2. Total number of downloads by each Operating System (group similar ones).
3. Top 10 (distinct) largest sized packages.
4. What were the top 10 least popular (distinct) packages?
5. At what specific hour there are most of the download hits?
6. What are the 5 most popular packages in US?
7. Show all packages downloaded by the machine with highest number of downloads?
8. Show top three OSs that are most popular among the R programmers?
9. How many R users still use 32 bit machines?
10. Show total number of downloads by each country, use ascending order?

Submission

- Submit your solution on Moodle before the deadline.
- Acceptable file format: Python notebook - name it `assignment2.ipynb`. The notebook should be exported as iPython Notebook with *.ipynb extension. If the code in your notebook does not run, it will result in 20% penalty.
- Take two screenshots of your solution to each question (code + its output into the Cassandra table) and insert it in a word document, generate a pdf of this document.
- Zip both files together and submit your solution on Moodle by the deadline.
- Do not submit work thats not your own and do not let others copy work that is your own. Both Copyier and Copyee will get ZERO marks.