**Lab Assignment #3 –**Spark standalone analytics

Due Date: Friday, Week 7

Purpose:

The purpose of this Lab assignment is to:

1. To carry out some basic pre-processing steps to prepare a big data environment for machine learning.
2. To convert native types to Spark types.

General Instructions:

Be sure to read the following general instructions carefully:

1. This assignment must be completed individually by all the students.
2. Only provide the requested screenshots and make sure to have a complete screenshot, partial screenshots will not earn any marks.
3. You will have to add all the analysis and screenshots in the Analysis report.
4. You will have to provide a **demonstration video for your solution**and upload the video together with the solution on **eCentennial** through the assignment link. See the **video recording instructions** at the end of this document.
5. In your 5-minute demonstration video you should explain your solution clearly, going over the main code blocks and the purpose of each method also demoing the execution of the instructions. Youtube links and links to google drive or any other media are not acceptable, the actual recording file must be submitted.
6. Any submission without an accompanying video will lose 25% of the grade.
7. Any submission without an accompanying Analysis report will lose 20% of the grade.

Assignment Pre-requisites:

1. Spark
2. Datasets indicated in the exercises
3. VMware platform image for COMP 251

Assignment Exercise

1-Load two days of data from the retail sales data available on the VMware image under the directory /home/centos/data/retail-data/by-day/ into a dataframe and name it df\_x where x is your firstname.

(use infer schema)

For students with firstname starting A-M please load the data for ninth and the tenth of December

For students with firstname starting M-Z please load the data for ninth and the tenth of January

2- Check the UI at the local host port 4040 or the port that spark connects to when launched and record the following in your analysis report:

The time it took to load the data.

The number of tasks and try to explain in your own words what happened in the analysis report.

Take a screenshot of the DAG execution and add it to your analysis report.

For points 3&4,5,6&7 use the Dataframe high level API, make sure you show the full column content, i.e. no truncation.

3- Carry out some basic investigation: count the number of records, print the inferred schema. Record the results in your analysis report.

4- Show all the transactions that are related to the purchase of stock id that starts with "227" with the type of product “ALARM CLOCK” mentioned as part of the description or a unit price greater than 5.

5- Store the results into a new dataframe name it df2\_firstname.

6- Show the sum of the quantities ordered and the minimum quantity order and the maximum quantity order for the transactions you extracted in point 4 above. Investigate the UI take a screenshot of the DAG plan and in your” Analysis report” add the number of stages the job required with the total time required per stage in addition to the number of tasks required for each job. Finally, drill down on each stage and produce the DAG graph for each stage and analyze the statistics, note the shuffle size and the number of partitions in your report.

7- Show all the transactions mentioned in point 4 above that have originated form outside the United Kingdom.

**Naming and Submission Rules:**

1. You must name your submission according to the following rule:

**YourFullname\_COMP251assignmentnumber**.Example: **AdamPerjouski\_COMP251assignment1**

1. Please add all the commands/instructions into a python script.
2. There is no need to zip the files.
3. Upload the submission file on e-Centennial using the Assignment link(s).

**Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation criteria | **Not acceptable** | **Below**  **Average** | **Average** | **Competent** | **Excellent** |
|  | **0% - 24%** | **25%-49%** | **50-69%** | **70%-83%** | **84%-100%** |
| Requirements in exercises  50% | Missing all requirements required | Some requirements are implemented. | Majority of requirements are implemented but some are malfunctioning. | Majority of requirements implemented. | All requirements are implemented  Correctly. |
| Instruction/ Code Documentation on python script  5% | No comments explaining code.  Missing screenshots | Minor comments are implemented. | Some code is correctly commented. | Majority of code is correctly commented. | All code is correctly commented. |
| Written analysis  Content  15% | Missed all the key ideas; very shallow. | Shows some thinking and reasoning but most ideas are underdeveloped. | Indicates thinking and reasoning applied with original thought on a few ideas. | Indicates original thinking and develops ideas with sufficient and firm evidence. | Indicates synthesis of ideas, in-depth analysis and evidences original thought and support for the topic. |
| Written analysis report format and organization  5% | Writing lacks logical organization. It shows no coherence and ideas lack unity. Serious errors. No transitions.  Format is very messy. | Writing lacks logical organization. It shows some coherence but ideas lack unity. Serious errors.  Format needs attention, some major errors. | Writing is coherent and logically organized. Some points remain misplaced.  Format is neat but has some assembly errors. | Writing is coherent and logically organized with transitions used between ideas and paragraphs to create coherence. Overall unity of ideas is present. Format is neat and correctly assembled. | Writing shows high degree of attention to logic and reasoning of all points. Unity clearly leads the reader to the conclusion.  Format is neat and correctly assembled with professional look. |
| Demonstration Video  25% | Very weak no mention of the code changes. Execution of code not demonstrated. | Some parts of the code changes presented.  Execution of code partially demonstrated. | All code changes presented but without explanation why. Code demonstrated. | All code changes presented with explanation, exceeding time limit. Code demonstrated. | A comprehensive view of all code changes presented with explanation, within time limit. Code demonstrated. |

**Demonstration Video Recording**

Please record a short video (max 4-5 minutes) to explain/demonstrate your assignment solution. You may use the Windows 10 Game bar to do the recording:

1. Press the Windows key + G at the same time to open the Game Bar dialog.

2. Check the "Yes, this is a game" checkbox to load the Game Bar.

3. Click on the Start Recording button (or Win + Alt + R) to begin capturing the video.

4. Stop the recording by clicking on the red recording bar that will be on the top right of the program window.

(If it disappears on you, press Win + G again to bring the Game Bar back.)

You'll find your recorded video (MP4 file), under the Videos folder in a subfolder called Captures.

Or

You can use any other video recording package freely available.