**Lab Assignment #1 – Data lake**

Due Date: Friday, Week 2

**Purpose:**

The purpose of this Lab assignment is to:

1. To get introduced querying big data.
2. To get hands-on experience of setting up a data lake on the cloud.
3. To carry out some basic analytics on a cloud data lake.

Throughout the assignment we will carry out the following activities:

1. Create and configuring the Data Analytics on Azure (Exercise #1 in module #1 lab tutorials)
2. Query the data lake

**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 accompany your solution submission with an analysis report that contains your findings and the required screenshots.
4. You will have to provide a **demonstration video for your solution related to exercises 2 & 3** 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. Youtube links and links to google drive or any other media are not acceptable, the actual recording must be submitted.

**Assignment Pre-requisites Instructions:**

1. A Microsoft azure free student account
2. Synapse analytics workspace and azure Data Lake Gen2 set up as per the instructions in exercise #1 Lab tutorials module #1

**Exercises**

Exercise #1: Research the parquet format (15 marks)

In the lab tutorial of module #1 we loaded a parquet file. Research this type of file, and provide a one-page summary describing the following:

1. A bit of history of when it was introduced and what are the main benefits of this format.
2. The main structure and technologies behind it.
3. Which software/platform (s) support it.
4. A comparison between csv and parquet.
5. List two other file types comparable to parquet.

Be thorough in your research and make sure you use your own word and reference all your sources at the end of the analysis report.

Exercise #2: setup your workspace (20 marks)

1. Log into your azure portal and expand your data lake resource group created during lab exercise #1 and take a screen shot showing both the storage account and the analytics workspace. Make it contains all the naming explained in lab #1.
2. Expand the storage account view and take a screenshot showing the first page and add it to your analysis report.
3. Click on the “JSON view” tab, examine the output. Copy the content of the JSON file and add it to your analysis report.
4. On the left side click on the “Containers” tab under “Data storage”, take a screenshot and add it to your analysis report.
5. Navigate back to the resource group and expand the “Synapse workspace” , take two screenshots showing all the configurations and add that to your analysis report.

Exercise #3: Retrieve data using a synapse serverless SQL pool (35 mark)

1. On Microsoft Azure portal click/expand your “Synapse workspace” .
2. Click on the Workspace web URL to launch your “Synapse workspace” in another tab.
3. Using the develop hub, write an SQL script to access the pandamic datalake csv file, available at the following link <https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/latest/ecdc_cases.csv> In the statement you should retrieve the following information:
   1. The top 15 rows
   2. Only the following two columns:
      1. the column reflecting the total # of cases.
      2. The column reflecting the total # of deaths. (Hint: check the usage of the WITH )

Save the results as a row set named rowfirst name, where first name is your firstname.

1. Take three screenshots and add them to your analysis report, the first showing the query, the second showing the results and the third showing the message.
2. In your workspace navigate to the monitor hub, click on the SQL requests, take a full screenshot and add it to your report.

**Evaluation**

|  |  |
| --- | --- |
| Solution and scripts are correctly completed. | 40% |
| Proper setup of resources on the cloud and execution of scripts. | 20% |
| Analysis report contains valid analysis and requested comparisons and screenshots. | 15% |
| Explanation in the video recording demonstrates the execution of the assignment exercises. | 15% |

**Naming and Submission Rules:**

1. Submit the mp4 video together with your solution and written response, the written response can be a word document or pdf document showing all the analysis & screenshots in the word document, in a zipped file.
2. You must name your submission according to the following rule:

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

1. Upload the submission file on eCentennial using the Assignment link. (One zipped file)

**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 may use any other recording service.

References

1. [Microsoft Synapse](https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-create-workspace)
2. [Openrowset](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset)
3. [serverless SQL pool](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/on-demand-workspace-overview)