DBS Assessment Brief

Assessment details

Unit Title	Tools for Data Analytics
Unit Code	B8IT106
Level	8
Assessment Title	Spark, Tableau
Assessment Number	2
Assessment Type	Individual
Assessment Weighting	20%
Issue Date	The 6th of November 2018
Hand in Date	The 26th of November 2018
Mode of Submission	On-line (Moodle)

Assessment Task (20%)

Do all of the below parts:

Part 1 (Spark)

a) Install PySpark. You can use your notes from the class or the following guides:

- for Linux:

https://www.youtube.com/watch?v=PRzSWWsyHZg&list=PL9ooVrP1hQOEBF5zdCdoMs2l1wws6be2X (from 4:30)

https://blog.sicara.com/get-started-pyspark-jupyter-guide-tutorial-ae2fe84f594f

- for Windows:

https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c

Clue: Please read/watch the guides for Linux as many steps are similar for Windows/Mac.

You will need:

- Java

- Python

- pyspark (downloaded from the spark webpage)

- setup environmental variables

- you can – and it is recommended – to link pyspark with Jupyter Notebooks

Describe the difficulties you were facing during the installation and configuration.

b) Run spark notebook from the class (section Dataset API only) and modify it to conduct the analysis on a different dataset you like. At the beginning describe the dataset and the goal of your analysis. Write at least 10 different queries / commands using sql or dataframe APIs. Explain what each command / query does. After the analysis describe the insights.

   The comments and descriptions you can put directly in the notebook.

   Output : jupyter notebook or python file with comments.

c) Optional: you can also try to configure and execute RDD API part. If you implement 3 examples of transformation-action using RDD API you can get extra points.

   Clue: probably you will get an error which will be solvable with this:

   https://stackoverflow.com/a/50399085

Part 2 (Tableau)

a) Register for a free trial on Tableau web page. Watch Tableau introductory video:

   https://www.youtube.com/watch?v=GkJwcyI_1vc

   and follow the steps in the tutorial.

b) Load into Tableau the same dataset you were using for Part 1 of this CA. Conduct visual analysis (i.e. generate charts to learn about the dataset). In half a page describe your experience with Tableau by comparing it to Spark and Python. List Tableau differences, advantages and disadvantages versus Spark and Python.

   Output: Half a page text. You can include graphics from your analysis.

A zipped file should be submitted online (26th of November 23:30) including both an individual report and all supporting processes/workbooks/scripts/spreadsheets. The report should detail the steps followed and include relevant screenshots, challenges encountered and findings. The grade assessment will be based on the DBS CA grading scheme.