University of South Wales
Prifysgol De Cymru

Faculty Of Computing and Engineering Sciences

Assessment Cover Sheet and Feedback Form 2020-21

Module Code:	Module Title:	Module Team:
MS4S21	Big Data Engineering and Applications	Moizzah Asif
Assessment Title and Tasks: Set Tasks - not-time constrained 2		Assessment No. 1
Date Set: 28-Apr-2021 17:00	Submission Date: 14-May-2021 21:00	Return Date: 09-Jun-2021 21:00

IT IS YOUR RESPONSIBILITY TO KEEP RECORDS OF ALL WORK SUBMITTED

Marking and Assessment
This assignment will be marked out of 100% This assignment contributes to 60% of the total module marks.
Learning Outcomes to be assessed (as specified in the validated module descriptor https://icis.southwales.ac.uk/): 1) To appraise and contrast strategies for dealing with Big Data 2) To demonstrate an ability to apply Big Data concepts in non-trivial contexts
Provisional mark only: subject to change and / or confirmation by the Assessment Board

MS4S21 Coursework - I & II: 2020/21

Submission deadline: 14th May 2020 - 21:00

Your assessment is divided into two main parts: **Coursework I** and **Coursework II**

Coursework I is based on conducting experiments and providing the scripts along with brief overview and commentary in the sequence of execution in a single pdf file. It contributes 60% to this module's assessment.

Coursework II contributes 40% to the this module's assessment and is based on writing reports on specific big data technology research topics and tools.

You are required to:

1. submit a pdf file for both courseworks respectively,
2. submit any programming scripts and code used in Coursework I, and
3. attempt all the questions in both the courseworks.

You may attach appendices to each coursework's pdf file.

You should be able to complete these courseworks while staying well under your AWS classroom budgets. It will be your responsibility to not to exceed the budget and have contingency plans to evidence your work in the unlikely event that you loose your work due to reaching the budget limit. Furthermore, you are advised to plan the execution of your experiments for Coursework I in such a way that you are able to terminate any unused AWS services without affecting your coursework completion and progress.

The pdf and any accompanying files for both the courseworks should be submitted as separate submissions, under their own BB assignments provided in Blackboard on MS4S21 home page by the deadline.

MS4S21 Coursework - I: 2020/21

Submission deadline: 14th May 2021 - 21:00

Contribution to module: 60%

This coursework is divided into three main Experiments: **Experiment I, II** & **Experiment III**.

You are required to conduct all experiments and provide relevant Linux-terminal, python, and any AWS CLI scripts, and evidence to the approach to execute tasks on AWS's GUI in the order of execution as a single pdf file. You may add any other appropriate evidence of work that you may have done outside of the resources mentioned here.

Any work done on Jupyter notebooks, or any other Python IDE should be provided in ipynb or .py file formats respectively.

Please ensure that scripts for all experiments are placed under appropriate sections with relevant and intuitive names.

You should also provide supplementary overview and commentary to reflect your approach as you evidence your work for each experiment.

Experiment I

10 marks

You are required to attempt this task on the hadoop cluster created using Ubuntu VMs on local machines during the lectures.

1. Replace the existing worker node with a new worker node with following characteristics and hardware specifications:

    - 14 Gb of RAM;
    - 2 processors/cores/CPUs;
    - should run on a Linux version which is consistent with master node's OS;
    - name the new VM as 'worker-cw';
    - should have two user accounts, out of which one of the user accounts should be named 'hadoop-cw'; and
    - any hadoop communication and functionality on worker-cw node should be carried out only via hadoop-cw user account.

    Your cluster should have 2 nodes including the existing master node and the worker node that you are required to create in this coursework. You should be able to demonstrate that your newly configured hadoop cluster is ready to execute a map reduce job.

Please provide the following in your pdf file.

(a) A sequential script that can be executed to replicate your experiment. The script should consist of all the terminal commands from your hadoop nodes (i.e. linux VMs), in the sequence of execution. Include any debugging/troubleshooting commands that you may require to execute as part of the same sequential script and in order of execution.

(b) A copy of configuration tags for each of the hadoop configuration files that you may have edited. Paste the configuration tag after the relevant text editing Linux-terminal command in your script;

for example if you have opened the start-dfs.sh file using vi editor and made some edits; you should copy and paste those edits after the terminal command which you used to open this shell script.

(c) The modifications made to bash file/s. (please don't copy and paste the predefined scripts and variables from bash file). Use the same strategy to present your work as suggested in (b).

(d) A copy of any system files that you may have edited on Linux VMs. Use the same strategy to present your work as suggested in (b).

(e) Screenshots of inputs, outputs on terminal and any text editors that you may use.

Experiment II

30 marks

1. Stream public tweets from twitter which are authored in the United Kingdom over a period of two days. The tweets must be: in English language, and contain either of the following hashtags or word: **#COVID19 #CoronaVirus #UnitedKingdom #UK** and **#India**. The twitter stream should span across exact two days, i.e. 48 hours. You can either use and modify the tweets streaming script provided in lecture or create your own.
Note: Please mention the dates with start and end time in comments when you download the tweets.

2. Save only the relevant fields from the streaming API tweet response in a csv file so that you can perform the next task and Experiment III.

You are also required to pre-process any data at this stage preferably by using python programming language. The pre-processing would retain the appropriate attributes from response fields as features/CSV-columns and must include creation of new features as follows:

- a feature called tweet_text where all English language stop words, any punctuation marks and hashtags/words mentioned in 1 have been removed from the tweet's text; and

- the pre-processing should also include creation of 6 other features each representing one of the hashtags/words from 1. The value of these feature would be the frequency of times the hashtag/word has appeared in the tweet.

You should save the final set of features in a csv file.

3. Create a database in DynamoDB using the CSV file that you have streamed in Task 1. Your database should be modelled in a way that comparison between the progress of COVID-19 and any lockdown applied or lifted in India and UK can be drawn using results from simple NoSQL style queries. Note that you are not required to run any queries to do analysis at this stage. You must demonstrate the following in your solution:

4. importing data from the CSV file to DynamoDB,

5. a data model to reflect your database design, and

6. any python script and code used, as well as screenshots of actions performed on AWS GUI to perform this task.

Experiment III

20 marks

You are required to run HiveQL queries on the tables that you have created in Experiment II to find the following:

1. the country which has authored the most tweets,

2. the most frequent hashtag/word mentioned in Experiment II found in tweets from each country,

3. the most frequent hashtag/word mentioned in Experiment II found in all tweets,

4. total number of user mentions in tweets from each country respectively, and

5. total number of user mentions in all tweets.

Your approach to address this task should clearly indicate which big data cloud technologies were used and how was the data moved across and queried from one platform to another.