Cloud Computing Assignment

The students have the choice of solving either problem scenario 1 or 2.

Problem Scenario 1

The task at hand is the development of a collaborative platform for archival and sharing of class notes and student coursework.

The platform should allow handling of:

- Teacher accounts;
- Groups of two/three students;
- A repository of documents;
- Appropriate measures against data loss;
- Support for versioning of the stored documents;

The documents may be stored in source format (LaTeX, docx) or in PDF. Optionally, for source format submissions you can implement a solution to generate automatically the appropriate PDF file.

The collaboration site shall be implemented on the Amazon AWS platform.

It is suggested to use the DRUPAL software, `www.drupal.org`, for the deployment of the website. A tutorial on installing DRUPAL is available at `https://www.tecmint.com/install-drupal-in-centos-rhel-fedora/`. Two **important** additional install instructions.

- At the very start, edit the file `/etc/selinux/config` so that `SELINUX=disabled`.
- After drupal has started up and you have edited `/etc/httpd/conf/httpd.conf`, run the command `systemctl restart httpd`.

Using Drupal is *not mandatory*: you may implement the system using any tools you feel comfortable with.

You will have to turn in:

1. A report containing a description of the approach taken, the design of the solution, some examples of use, and a project and test plan;
2. The source code;
3. The web site operational and available for testing.

You will ensure that the website is operational for two days after the submission deadline and made available for testing by the module leader as part of the marking process; as part of the submission process you will send the module leader the necessary information to connect to the application, including creation of an account where appropriate. *You shall turn off the AMI instance hosting the application at other times whenever you are not working on it.* You will be informed after testing is done, at which point you will be able to stop the AMI instance. At a later time you will also be informed that the marking/course

administration is completed, at which point you will be able to terminate the AMI instance. You will also be notified when your AWS account will eventually be unlinked from the main Cranfield AWS account.

Very important: Do not use the Amazon RDS service. Whilst it is a perfectly sensible solution from a purely technical point of view, it has a high cost that is not justified in the context of a simple assignment exercise, and it would consume your AWS Educate credits in a very short time.

Problem Scenario 2

The task at hand is to use the Spark infrastructure on AWS to solve at least one of the queries for the DEBS 2015 Grand Challenge `https://debs.org/grand-challenges/2015/`. The dataset records a years worth of taxi trips for the city of New York. The website contains a link (Google Drive) to a subset of the data for testing purposes. Learn how to easily download Google Drive files at `https://stackoverflow.com/a/39225039`, following these instructions:

```
pip3 install gdown
gdown https://drive.google.com/uc?id=0B0TBL8JNn3JgTGNJTEJaQmFMbk0
sudo yum install gzip.x86_64     /* if gunzip is not installed */
gunzip sorted_data.csv.gz
```

The challenge proposes two queries:

Query 1 - Frequent Routes: Find the top 10 most frequent routes during a window covering the last 30 minutes;

Query 2 - Profitable areas: Find the top 10 most profitable areas during a window covering the last 30 minutes;

The specification "over the last 30 minutes" means that the solution is intended to be applied to a data stream, so that the system can be in principle applied to real-time monitoring of the network. For the submission requirement of this scenario, please see previous "You will have to turn in" section. For both queries it is assumed that the result is calculated in a streaming fashion – i.e.: (1) solutions must not make use of any pre-calculated information, such as indices and (2) result streams must be updated continuously. You are free to choose either one of the queries.

Further details can be found at the aforementioned link.

Marking Scheme

40 Software (& website for scenario 1);

40 Report document;

20 Test design and results

Submission deadline: 09:30 am, February 22nd, 2021.
1. Website and application for scenario 1 to be available from submission time until 6:00pm, February 26th, 2021.
2. Submission deadline for Part Time Students: 09:30 am, March 08th, 2021.