University Of Engineering and Technology, Lahore
Computer Engineering Department

Course Name: Database Systems	**Course Code: CS363L**
Assignment Type: Complex Engineering Activity	**Dated: 14-02-2022**
Semester: 6th	**Session: 2019**
Lab/Project/Assignment #: Project 2	**CLOs to be covered: CLO1, CLO4**
Lab Title: Visualizing Data	**Teacher Name: Ms. Darakhshan**

Complex Engineering Activity (CEA)

	Attribute	Complex Activities	Apply/Not Apply	PLOs Covered	Taxonomy Level
1	Preamble	Complex activities means (engineering) activities or projects that have some or all of the following characteristics listed below:	N/A		
2	Range of resources	Involve the use of diverse resources (and for this purpose, resources include people, money, equipment, materials, information, and technologies).	Applicable		
3	Level of interaction	Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering, or other issues.	Applicable		
4	Innovation	Involve creative use of engineering principles and research-based knowledge in novel ways.	N/A		
5	Consequences to society and the environment	Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation.	N/A		
6	Familiarity	Can extend beyond previous experiences by applying principles-based approaches.	Applicable		

CEA Description:

In this Complex Engineering Activity, students will learn how to visualize big datasets using colaboratory. They will understand BigQuery and Python integration and write optimized queries. They will also learn how to analyze the performance of a query in terms of how many bytes it requires to get processed. They will then analyze the visualizations to understand how certain features of a GitHub repository correlate to its popularity.

Rubrics

	Implementation
50	Completed all questions **(including bonus/Q6)** and understands how tasks were solved + Have written optimized Queries and have used visualization techniques to understand dataset. Understands Github dataset, Python, BigQuery and their integration. No plagiarism.
40	Completed all questions **(except bonus)** and understands how tasks were solved. Students have used visualization techniques to understand dataset and know how to do this in Python Worked upon optimized queries. Understands BigQuery and Python Integration Student understands Github dataset. No plagiarism.
30	Completed **Question 1, 2 and 3** and understands how tasks were solved. Student understand how to **visualize** datasets using Python and how to integrate BigQuery with Python. Student understands Github dataset and has worked on **writing optimized queries**. No plagiarism.
20	Completed **Question I and 2 question only** + Integration of **BigQuery with Python** is complete + **Understands Query Efficiency**. Student understands Github dataset. No plagiarism.
10	Solved **Question I only** and **understands** the **dataset**. No plagiarism.
0	Project missed or solved none of the problems

Introduction:

In this project you will explore a public GitHub dataset using Colaboratory. Colaboratory is like a Jupyter notebook, but it has collaboration and integrations with BigQuery built into it. We will be using a subset of the BigQuery public dataset due to its size.

You will be exploring the dataset in the provided notebook project2.ipynb. You can access the notebook from the course website, **make a copy of it in your own drive**, and begin the assignment. Note that this part is intended to help prepare you for the last course project, where you will be running through an entire data cycle of querying, visualizing, and predicting on a dataset of your choosing. This is an individual project, but getting used to using Colab will aid you in Project 3 when you can work in pairs.

How to Get Started:

Get Setup with Colaboratory

Here is an overview of Colaboratory features and a brief guide for using BigQuery through Colaboratory. Before proceeding, make sure you have read and understood these support documents. To open a new notebook in Colab, you can go to File > Upload notebook and choose the file either from your computer or from Google Drive. You can also make a copy of an existing Colab notebook by going to File > Save a Copy in Drive. Colab notebooks can be saved just like any other file to your own Google Drive account.

Section 1: Understanding the GitHub Dataset (4 points):
To begin your exploration of the GitHub dataset, you will briefly investigate the various tables in the dataset and answer a few short questions. This will help you understand what tables to use for your visualizations in Section 2.

Section 2: Investigating Query Performance (8 points):
In this task, you will look at some queries that are pretty inefficient in terms of how many bytes need to be processed. You will think about what makes queries inefficient and how to make them more efficient. You'll also get a chance to look at how the query is optimized in BigQuery.

Section 3: Visualizing GitHub Data (38 points):
This task will be the bulk of Project 2. You will learn to create visualizations to help you understand and answer questions about the GitHub data. For this assignment, you will have to think about what data you should use to answer a question and what kind of visualization you should make to clearly convey your information. You will then analyze your visualizations to understand how certain features of a GitHub repository correlate to its popularity.
This part of the project will be valuable practice for Project 3.

Honor Code:
This assignment is to be done individually. We encourage students to form study groups to complete the assignment, but the solutions to each assignment must be written independently. Be sure to list your collaborators on each part of the assignment at the top of the corresponding Colaboratory notebook.

We take the Honor Code seriously. Working in groups to discuss class concepts or a specific problem at a high level is OK, but the following would be considered honor code violations:

- Looking at the writeup or code of another student.
- Showing your writeup or code to another student.
- Discussing a problem in such detail that your solution is almost identical to another student's solution.
- Uploading your writeup or code to a public repository (where other students may be able to find it).

Submission Guidelines:

Once you have filled out the Colab notebook completely, you are ready to submit.

In total, you will be submitting two files separately. To submit:

1. Download the Colab notebook as an iPython notebook - you can do this by going to File > Download .ipynb.
2. Create a PDF of your Colab notebook, **making sure that you have run all cells first.** Make sure you've closed the table of contents sidebar before you create the PDF so we can easily see your work and output.
    - In Google Chrome, you can do this by going to File > Print and then choosing "Save to PDF". If your graphs are too large to fit in the PDF, you can try going to More Settings, setting "Scale" to "Custom", and choosing a smaller scale (try between 40 and 50).

3. Submit the PDF file to the **2019-CE-X_project2 _PDF** assignment on Google classroom.
4. Submit the iPython notebook to the **2019-CE-X_project2_submission** assignment on Google classroom by **Sunday, 6th March, 2022 9 P.M**.

Note: We reserve the right to deduct points from your project if you do not follow the submission instructions, if there are some cells which have not been run or with non-readable output, or if you have assigned your pages to the questions on Grade scope incorrectly. **Please read through your PDF document before you submit it and *ensure that all answers are clearly visible.*** Please also leave yourself enough time to do the assignment/submission, and go over your assignment to make sure it is correct!