CSP 554–Big Data Technologies

Assignment 01 (Modules 01a & 01b, 50 points)

The assignment is due before midnight for the next class period.

Assignments are to be uploaded via the **Canvas** portal.

1. Obtain our texts
   - **Tom White. 2015. *Hadoop: The Definitive Guide* (4th ed.). O'Reilly Media, Inc (TW)**
   - **Pramod J. Sadalage and Martin Fowler. 2012. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley. (PS)**

2. Read from (TW)
   - **Chapter 1 (note this chapter is also on Canvas "Free Books and Chapters," so you do not need to wait for the book to arrive)**
   - **Chapter 3**

3. (**5 points**) Read an article on "**Canvas**":
   The Parable of Google Flu (3 pages!)
   Summarize the article in 1-page and not more.

4. **(30 points**) Read an article on "Canvas":

   A brief survey on big data: technologies, terminologies and data-intensive applications (36 pages)

   Summarize the article in 2-page and not more.

(Please refer to "**How to Summarize a Research Article**")

5. **(5 points)** Answer each of the following questions about the article in just one to three sentences each:
   - What was the problem with the Google flu detection algorithm?
   - What is big data hubris?
   - What approach could have been used to improve the Google flu detection algorithm?
   - What is "algorithm dynamics?"

- What aspect of algorithm dynamics impacted the Google flu detection algorithm?

6. **(5 points)** Set up a Google Cloud platform (GCP) cloud account if you do not already have one (see below for details). This will get you started since we will do most of our GCP assignments.

   a. We will be using GCP services. So, if you do not have access to a regular personal (or business) GCP account, you must create one (we will point you to step-by-step instructions describing how to do this). To establish a regular account, you will need a valid credit card. Assignments in this course are carefully structured, so your cloud costs should be minimal (no more than $5-$7 per month, and usually much less). But you must follow assignment directions and remember to release cloud resources when instructed; otherwise, your costs may be (much) higher.

   b. To sign up for an GCP account, you need to follow the first two modules of the "Setting Up Your GCP Environment" tutorial located at the link below (but be sure to read the notes below before you start):

      https://console.cloud.google.com/freetrial/signup/

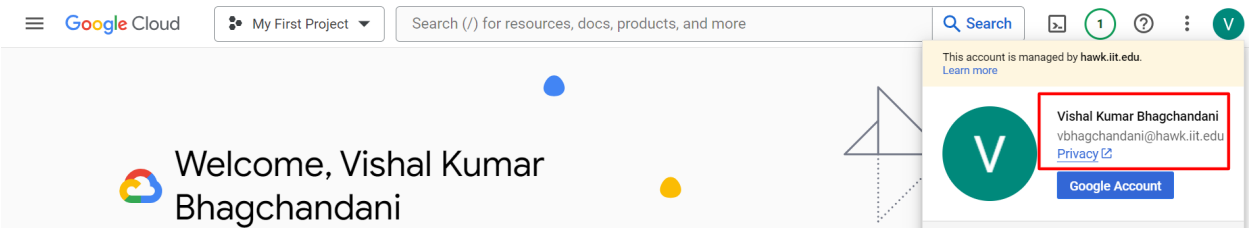
      You need to complete the two modules: "Create your GCP account" and "Configure users".

      Note 1: The email that GCP sends to verify your email address often appears in your spam folder, especially if you use Gmail. So, wait a few minutes and then look in all your email folders.

      Note 2: If you can make it all the way through the module "Create your GCP account" then you can perform all course exercises. Successfully completing the module "Configure users" is important, if you can do it, as it will ensure your GCP account is more secure from malicious access. Next best, is completing the module "Configure users" at least to the point where you have set up multi-factor authentication (MFA) on your root account.

      Note 4: At some point when following the "Configure users" module you will need to choose and authenticator app for your mobile phone. Microsoft Authenticator is a good free choice. If you install this app, when the time comes scan for a QR code if you are using an iPhone (a) press "+" in the top right corner of the app, and (b) select "Other (Google, Facebook etc.)", then scan for the QR code. It should be similar for an Android phone. But if you prefer Google Authenticator, or some other app, that is fine too.

c. To get credit for this part of the assignment, provide a screen shot of the main page of the GCP management console page including your account name (see arrow)similar to the example below:



7. (**5 points**) Now follow the below steps about how to work with an GCP cloud storage service called GCS (Google Cloud storage). In a while we will come to understand GCS as one critical element of a big data processing architecture known as the "data lake.

   a. GCS (Google Cloud storage) is storage for the internet. You can use GCS (Google Cloud storage) to store and retrieve any amount of data at any time, from anywhere on the web. Google GCS stores data as objects within buckets. An object is a file and any optional metadata that describes the file. To store a file in GCS, you upload it to a bucket. When you upload a file as an object, you can set permissions on the object and any metadata. Buckets are containers for objects. You can have one or more buckets. You can control access for each bucket, deciding who can create, delete, and list objects in it. You can also choose the geographical Region where GCS will store the bucket and its contents and view access logs for the bucket and its objects.

   b. If you want more details about GCS storage refer to the section "Overview of GCS" at the end of this document

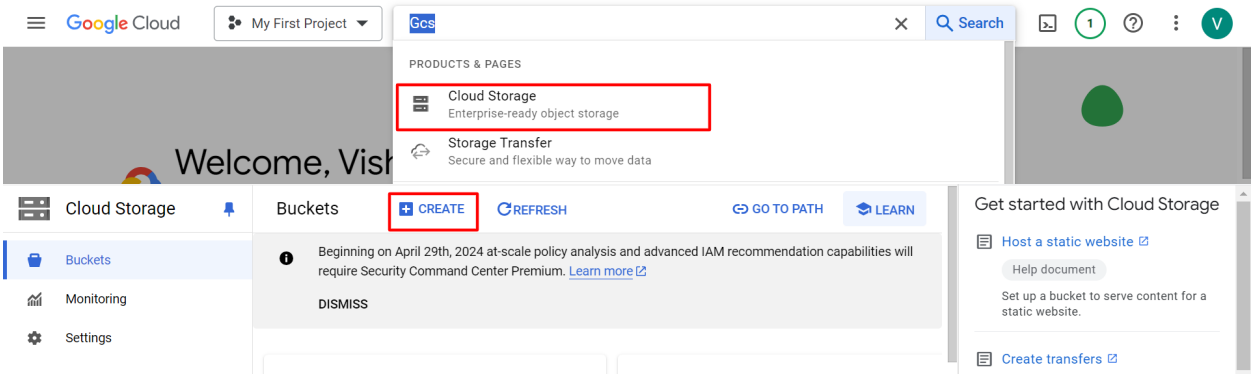
   c. Creating a bucket

      Now that you have signed up for GCP, you are ready to create a bucket using the GCP Management Console. Every object in Google GCS is stored in a bucket. Before you can store data in Google GCS, you must create a bucket.

      Note: You are not charged for creating a bucket. You are charged only for storing objects in the bucket and for transferring objects in and out of the bucket.

      To create a bucket:

      - Sign in to your GCP account. Type "GCS" into the services search box and then press the "Enter" key:



- Notice the menu on the left side of the page. Each entry, such as "Buckets" results in the display of a different screen, which allows you to perform specific functions. Sometimes the menu collapses to save space. It is then replaced by three stacked horizontal bars in the top right corner of the screen. Click on that symbol to reveal the full menu.
- Choose the "Create bucket" button towards the top middle of your screen and then the Create bucket page opens.
- In Bucket name, enter a DNS-compliant name for your bucket. The bucket name must: (a) Be unique across all of Google GCS bucket names in the world; (b) Be between 3 and 63 characters long; (c) Not contain uppercase characters; (d) Start with a lowercase letter or number After you create the bucket, you can't change its name. As one possibility, the bucket should be named something like "YourIITId-CSP554", for example: "A1234567_CSP554"
- In Region, choose the GCP Region where you want the bucket to reside. Choose a Region close to you to minimize latency and costs and address regulatory requirements. Objects stored in a Region never leave that Region unless you explicitly transfer them to another Region.
- Scroll down to see more of the screen.
- Keep scrolling down until you see the "Create bucket" button. Choose Create bucket. Now you've created a bucket in Google GCS

d. Now that you've created a bucket, you're ready to upload an object to it from your PC or Mac (or any other computer). An object can be any kind of file: a text file, a photo, a video, and so on.

   To upload an object to a bucket

- In the Buckets list, choose the name of the bucket that you want to upload your object to. To do so, just click on the name of your bucket, and then the Objects screen should appear.
- On the Objects screen for your bucket, choose Upload.
- Under Files and folders, choose Add files.
- Choose a file to upload, and then choose Open. When asked to upload a file to the GCS bucket you have created, just use any text file you have handy (even this one).
- Choose Upload (this is a button towards the bottom of the page). You've successfully uploaded an object to your bucket.

To receive credit for this question, provide a screen shot showing some named object in the bucket. Note, this is the screen that appears after you choose the Upload button.

e. If you completed this assignment and do not plan to use your bucket or objects, we strongly recommend that you delete your bucket so that charges no longer accrue. Before you delete your bucket, you must empty the bucket or delete the objects in the bucket. After you delete your objects and bucket, they are no longer available.

   To empty a bucket
   - In the Buckets list, select the bucket that you want to empty (by clicking on the little circle to the left of the bucket name), and then choose Empty.
   - To confirm that you want to empty the bucket and delete all the objects in it, in Empty bucket, enter the words "permanently delete."
   - Important, emptying the bucket cannot be undone. Objects added to the bucket while the empty bucket action is in progress will be deleted.

   Important: Deleting a bucket cannot be undone. Bucket names are unique. If you delete your bucket, another GCP user can use the name. If you want to continue to use the same bucket name, do not delete your bucket. Instead, empty and keep the bucket.

   Deleting your bucket

   - From the left-hand side menu select "Buckets" and the list of buckets should appear
   - To delete a bucket, in the Buckets list, select the bucket (by clicking on the little circle to the left of the bucket name).
   - Choose Delete.
   - To confirm deletion, in Delete bucket, enter the name of the bucket.
   - Now select the "Delete bucket" button

You are now done with this assignment!

Overview of GCS

Google Cloud Storage is a scalable and secure object storage service that allows you to store and retrieve any amount of data. It's designed to serve a variety of use cases, including live data serving, analytics, and machine learning.

Advantages of GCS

Google Cloud Storage offers several advantages that make it a robust and versatile choice for data storage:

- **Create Buckets**: You can create and name a bucket to store your data. Buckets are the primary containers in Google Cloud Storage.
- **Store Data in Buckets**: Store an unlimited amount of data in a bucket. You can upload objects up to 5 TB each, and each object is uniquely identified by a key.
- **Download Data**: You can download your data at any time, or allow others to do so by sharing access.
- **Permissions**: Control who can upload or download data from your buckets. You can set permissions for individual users or groups.
- **Standard Interfaces**: Google Cloud Storage supports RESTful APIs and provides client libraries for various programming languages, making it easy to integrate with other tools and services.

Google GCS Concepts

This section describes key concepts and terminology you need to understand to use Google GCS effectively. They are presented in the order you will most likely encounter them.

Buckets

A bucket is a container for storing objects in Google Cloud Storage. Every object resides in a bucket, and buckets serve several purposes:

- **Organization**: Buckets help organize the storage namespace.
- **Billing**: They determine the account responsible for storage and data transfer charges.
- **Access Control**: Buckets manage access permissions.
- **Location**: You can create buckets in specific geographic locations to optimize latency and availability.

For example, an object named **photos/puppy.jpg** stored in the bucket **my-photos** would be accessible at the URL **https://storage.googleapis.com/my-photos/photos/puppy.jpg**.

Objects

Objects are the basic entities stored in Google Cloud Storage. Each object consists of data and metadata:

- **Data**: The content of the object.
- **Metadata**: Key-value pairs that describe the object, such as the date it was last modified and its content type.

Objects are uniquely identified within a bucket by a key, which is essentially the object's name.GCP

Keys

A key is the unique identifier for an object within a bucket. Each object has one key, and the combination of the bucket name and key uniquely identifies the object. For instance, the URL **https://storage.googleapis.com/my-bucket/my-object** identifies the object **my-object** in the bucket **my-bucket**. Understanding these concepts helps you efficiently store, manage, and retrieve data using Google Cloud Storage.