CSP 554—Big Data Technologies

Assignment #1 (Modules 01a & 01b, 12 points)

Due by the start of the next class period

Assignments are to be uploaded via the Blackboard portal.

Note: There may be short quiz questions about the readings or articles and other questions available around the time of the second lecture.

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 Blackboard "Free Books and Chapters" so you don't need to wait for the book to arrive)
   - Chapter 3

3. (2 points) Submit very brief answers (or bullet points) to the following questions:
   - What location or time zone are you in when you attend the course?
   - Describe any prior experience you might have with use of public cloud, data mining, machine learning, statistics, data science and big data.
   - Share any big data interests and personal learning goals for the course.
   - Indicate if there are additional topics in the scope of the course of special interest to you.
   - Do you have any anticipated personal issues such as expected absences or other necessary accommodations with course impact? (Of course, these will be held in strictest confidence.)

4. Read article on "Blackboard" in Articles section
   - The Parable of Google Flu (just 3 pages!)

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 an Amazon Web Services (AWS) cloud account, if you don't already have one (see below for details), and then follow the tutorial about how to work with a storage service called S3. Since we will do most of our assignments using AWS, this will get you started. In a while we will come to understand S3 as one critical element of a big data processing architecture know as the "data lake."

   a. Most of what you need to accomplish is described in detail in the tutorial document called "AWS01.pdf". Skim the document, but don't start following the instructions outlined in the document until you have looked at the following "flow chart". Depending on your situation, you may need to skip one of more of the described steps.
   b. Follow this flow chart to decide how you need to follow the AWS01.pdf document's instructions.

If you already have a personal (regular) AWS account

Skip the step "Sign up for AWS"

If your account has an IAM user configured

Skip the step "Create an IAM User"
Skip the step "Sign in as an IAM User"
Log on as the IAM user
Continue with the rest of the steps

If your account does not have an IAM user

Option 1:

Note, if setting up an IAM users seems too complicated,
you can try option 2 below

Sign in as the root user
Follow the step "Create an IAM User"
Follow the step "Sign in as an IAM User"

Continue with the rest of the steps

Option 2: (a bit less secure)

Sign in as the root user
Skip the step "Create an IAM User"
Skip the step "Sign in as an IAM User"
Continue with the rest of the steps

If you do not have a personal (regular) AWS account

Option 1:

You could give this a try, but you might find this type of account
too limiting to do our assignments, or even not available

Sign up for an AWS Educate account *(see note about setting this up below)*
Sign in as the root user
Skip the step "Sign up for AWS"
Skip the step "Create an IAM User" *(Does not work for AWS Educate account)*
Skip the step "Sign in as an IAM User"
Continue with the rest of the steps

Option 2:

Recommended, option; your cost for using AWS for assignments should be low, less than US$10 per month

Follow the step "Sign up for AWS"

If you set up a standard AWS account for the first time DO NOT select a support plan. They are costly and you don't need one. See this URL for more details: https://aws.amazon.com/premiumsupport/knowledge-center/create-and-activate-aws-account/

Sign in as the root user

Note, if setting up an IAM users seems too complicated, you can skip the following two steps for now and move on as the root user

Follow the step "Create an IAM User"
Follow the step "Sign in as an IAM User"

Continue with the rest of the steps

c. Note on setting up an AWS Educate account:
   - If you can or prefer, you can set up an AWS educational account:
     - Step 1: Access the AWS Educate website https://aws.amazon.com/education/awseducate/ and click Apply Now.
     - Step 2: Click to Apply for AWS Educate for Students.
     - Step 3: Enter the information requested on the AWS Educate Student Application form.
     - Step 4: Verify your email address and complete a captcha to verify that you are not a robot.
     - Step 5: Click-through to accept AWS Educate Terms and Conditions.
     - After the application is submitted:
       - You will receive an email indicating that the application was received.
       - AWS Educate reviews the application and performs any necessary validation.
       - After you are accepted, a welcome message is forwarded to your email address. The message includes a link for the AWS Educate Student portal and an AWS credit code.

d. An overview of the S3 storage service is included in a section below for your reference.

e. To receive credit for this question, provide a screen shot showing the S3 bucket you have created. The bucket name should be named something like “YourIITId-CSP554”, for example: “A1234567_CSP554”

f. When asked to upload an object to the S3 bucket you have created, just use any text file you have handy (even this one).

g. Now also provide a screen shot showing some named object is in the bucket.

h. Make sure to follow the instructions in the pdf file for deleting your bucket at the end of the assignment so you do not incur additional costs.

Overview of Amazon S3

Amazon S3 a simple web services that you can use to store and retrieve any amount of data. It is used as part of a big data architecture called a “data lake” that we will discuss later.

Advantages to Amazon S3

Amazon S3 is intentionally built with a minimal feature set that focuses on simplicity and robustness. Following are some of advantages of the Amazon S3 service:

- Create Buckets – Create and name a bucket that stores data. Buckets are the fundamental container in Amazon S3 for data storage.
- Store data in Buckets – Store an infinite amount of data in a bucket. Upload as many objects as you like into an Amazon S3 bucket. Each object can contain up to 5 TB of data. Each object is stored and retrieved using a unique developer-assigned key. Data stored in S3 is “write once, read many.” This means that once written, data can’t be appended to or updated.
- Download data – Download your data or enable others to do so. Download your data any time you like or allow others to do the same.
- Permissions – Grant or deny access to others who want to upload or download data into your Amazon S3 bucket. Grant upload and download permissions to three types of users. Authentication mechanisms can help keep data secure from unauthorized access.
- Standard interfaces – Use standards-based REST interface designed to work with any Internet-development toolkit.

Amazon S3 Concepts

This section describes key concepts and terminology you need to understand to use Amazon S3 effectively. They are presented in the order you will most likely encounter them.

Buckets

A bucket is a container for objects stored in Amazon S3. Every object is contained in a bucket. For example, if the object named photos/puppy.jpg is stored in the johnsmith bucket, then it is addressable using the URL http://johnsmith.s3.amazonaws.com/photos/puppy.jpg

Buckets serve several purposes: they organize the Amazon S3 namespace at the highest level, they identify the account responsible for storage and data transfer charges, they play a role in access control, and they serve as the unit of aggregation for usage reporting.

You can configure buckets so that they are created in a specific region. For more information, see Buckets and Regions. You can also configure a bucket so that every time an object is added to it, Amazon S3 generates a unique version ID and assigns it to the object. For more information, see Versioning.

Objects

Objects are the fundamental entities stored in Amazon S3. Objects consist of object data and metadata. The data portion is opaque to Amazon S3. The metadata is a set of name-value pairs that describe the object. These include some default metadata, such as the date last modified, and standard HTTP metadata, such as Content-Type. You can also specify custom metadata at the time the object is stored.

An object is uniquely identified within a bucket by a key (name) and a version ID. For more information, see Keys and Versioning.

Keys

A key is the unique identifier for an object within a bucket. Every object in a bucket has exactly one key. Because the combination of a bucket, key, and version ID uniquely identify each object, Amazon S3 can be thought of as a basic data map between "bucket + key + version" and the object itself. Every object in Amazon S3 can be uniquely addressed through the combination of the web service endpoint, bucket name, key, and optionally, a version. For example, in the URL http://doc.s3.amazonaws.com/2006-03-01/AmazonS3.wsdl, "doc" is the name of the bucket and "2006-03-01/AmazonS3.wsdl" is the key.