* **Week 7 Assignment 1 - Securing AWS resources** *Please click on the link above to submit this week's assignment.*
* In this assignment you will create users, S3 bucket policy and modify the EC2 security group settings.
* User Creation
  1. First enable access to billing to other users by following Step 1 in this article.    http://docs.aws.amazon.com/IAM/latest/UserGuide/tutorial\_billing.html#tutorial-billing-step1
  2. Create 2 IAM users -  User1 and User2, and assign passwords
  3. Assign EC2 and S3 access to User1
  4. Assign only Billing and S3 access to User2
  5. Login as each user to confirm the settings
* S3 bucket policy
  1. Create an S3 bucket
  2. Create a bucket policy so that only User1 can read the contents
  3. Attach the bucket policy to the S3 bucket
  4. Login as User2 and check if it can or cannot access the bucket.
* Modify the EC2 security group settings to :
  1. Allow all Outbound traffic
  2. Only allow Inbound traffic for the following :  HTTP Port 80,   TCP Port 8080,  ICMP port,  SSH port 22
  3. NOTE: Start the EC2 server if its not running already
  4. Test the ping command to verify that the ICMP setting is correct.
  5. Test connecting to the EC2 server using SSH
  6. Now, go to the Security group setting and remove ICMP and SSH
  7. Test again with ping and ssh
* **NOTE: At the end of the assignment, don't forget to terminate the EC2 instance(s) and delete the S3 buckets and IAM accounts.**
* **Requirements for the written assignments:**
  1. This assignment and write up is due by end of Week 7
  2. Assignment file must have a .doc or .docx extension; screen shots should be in .jpg, .gif, or .pdf format
  3. Points for this assignment = 45


* **Week 7 Assignment 2 - Analyze use of compression and columnar storage types** Attached Files:
  1. [march.zip](https://lewisuniversity.blackboard.com/bbcswebdav/pid-3830117-dt-content-rid-35209940_1/xid-35209940_1) (1.22 MB)
  2. [jan.zip](https://lewisuniversity.blackboard.com/bbcswebdav/pid-3830117-dt-content-rid-35209941_1/xid-35209941_1) (1.142 MB)
  3. [feb.zip](https://lewisuniversity.blackboard.com/bbcswebdav/pid-3830117-dt-content-rid-35209942_1/xid-35209942_1) (1.081 MB)
* *Please click on the link above to submit this week's assignment.*
  1. Download the Airline On-Time performance data for the first quarter of any year to your PC, one month at a time. The description of the fields in the data is provided [here](http://www.transtats.bts.gov/Fields.asp?table_id=236). The files will be downloaded in zip format
     + Make sure that you choose the "dep\_del15" and "arr\_del15" fields since those will be needed later.
  2. Create the Hadoop cluster ( Bitnami, Cloudera, HDP or HDInsight)
  3. Using scp/WinSCP/Mobaxterm, copy the 3 files to the cluster
  4. Login to the cluster using ssh or putty
  5. Verify that the 3 files exist on the cluster
  6. unzip the 3 files one at a time.  After each unzip, rename the output file according to the month. For example,  January2019.csv and so on.
     + NOTE: if unzip is not available, install the unzip utility.
     + For Bitnami use,   sudo apt-get install -y unzip
     + For Centos use     sudo yum -y install unzip
  7. Combine all 3 files into one file using the following 2 steps :
     + First remove the header line from the February and March files, by using the 'sed' command
     + Join the January, February and March files together into one file using the Linux 'cat' command.
     + Verify the top and bottom of the combined file by using the 'head' and 'tail' commands
  8. Upload the combined data file to HDFS ( ex. to the /tmp directory)
  9. Verify that the file has been uploaded into HDFS
  10. If you are using Cloudera,HDP or HDInsight :  From a browser, go to the cluster's Hive view
  11. If you are using Bitnami, use the hive command line
  12. Create different tables with different names from the same data file stored in HDFS
      + Table-text stored as Textfile
      + Table-ord stored as ORC
      + Table-avro stored as Avro
      + Table-parquet stored as Parquet (if available)
      + NOTE:  When using Bitnami hadoop,  you must first create the tables first. Then load data into the textfile table. Finally, insert data into the other tables from the Text table.  (See instructions in the slides).
  13. From the Linux command window , run the HDFS commands to view the file sizes of the Hive tables
      + Go to the HDFS directory where Hive stores the files  ( example /hive/warehouse  or /user/hive/warehouse or something else...)
      + Run the HDFS command to display the sizes of the directories in the /hive/warehouse directory.  The command is "hdfs dfs -du -h /hive/warehouse.
  14. Start hive from the command line
  15. Display the list of databases and then list of tables
  16. Run the "describe formatted" hive command for each of the airline tables. Does the display type match the table type ( text, ORC etc.) ?
  17. **Optional :** Run the following queries for each of the tables
      + What is the total number of records ?
      + What is the total number of flights in the month of January?
      + What is the total number of flights with arrival and departure  delays over 15 minutes in each month? (Hint: see the dep\_del15 and arr\_del15 columns)
      + Write down the time it took to run each query for each type of table ( textfile, ORC, Parquet etc..)
      + Write down your observations about the timing of each query.
  18. Exit the browser and ssh/putty window
  19. Make sure that you have taken all screen shots for this assignment
  20. Don't forget to terminate the HDInsight cluster