***Only one attempt is allowed. You will not be able to resubmit this assignment.  So, review the assignment before submitting.***

1. Using the download instructions from Week 6 assignment, download the Airline On-Time performance data for the given month/year. Make sure that the size of your data files matches the values given in the table

|  |  |
| --- | --- |
| Months | April, May, June |
| Year | 2014 |
|  |  |
| Size of CSV files after unzipping | 56 MB |
| Total # of lines in the CSV files (with headers) | Enter your values here |

1. Transfer the files to an EC2 instance or to the VirtualBox VM.  Ensure that your VM/EC2 instance has sufficient disk space
2. Unzip the files one at a time.  After each unzip, rename the output file with your name, semester and the month-year. For example,  John-sp242-sept2021.csv and so on.
3. Combine all  files into one file using the following  steps :
   * Remove the header line from all files except the first file.  For example, remove from the 2nd and 3rd month files, by using the 'sed' command.
   * Join the 3 month 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
   * Use the **wc -l** command with the combined file to get the # of lines in the file.
   * **What** **is the total number of records in your dataset? Take a screenshot of the command.** 
     1. 
4. Use one of the following options (given below) to convert the data into various formats.
5. Check the file sizes after the conversion and provide your observations
6. NOTE: Rename the files, S3 bucket, AWS Glue resources etc.. with “sp242-004” either at the beginning or at the end of the name.
7. Provide the information in this table.

|  |  |
| --- | --- |
| Your Name | Simeon Ngalamou |
| Student ID | L30022598 |
|  |  |
| Which option did you use for this assignment? (Glue or Spark) | Spark |
| Month and Year of the data files | April 2014, May 2014, June 2014 |
| Number of lines in the combined CSV file | 1,485,395 lines |
| Size of the combined CSV file | 57 MB |
|  |  |
| ~~For Option #1 (AWS Glue)~~ |  |
| ~~Size of the~~ **~~Parquet~~** ~~file with~~ **~~snappy~~** ~~compression~~ |  |
| ~~Size of the~~ **~~ORC~~** ~~file with~~ **~~snappy~~** ~~compression~~ |  |
| ~~Size of the~~ **~~Parquet~~** ~~file with~~ **~~gzip~~** ~~compression~~ |  |
|  |  |
| For option #2 (Apache Spark) |  |
| Size of the **Parquet** file with **gzip** compression | 1.2MB |
| Size of the **ORC** file with **snappy** compression | 3.6MB |
| **Questions about the mystery file:** |  |
| Answer for question 8a | 19 |
| Answer for question 8b | Airline Code, Airline Description |
| Answer for question 8c | Apache Parquet |

**~~Option #1 :  Using AWS S3 and AWS Glue to convert data    
  
(NOTE: This option will result in a charge on your AWS bill.  Should be less than $5, but you need to monitor the charges)~~**

1. ~~Create a S3 bucket using the same name as you used in Week 4~~
2. ~~Upload the combined data file to an S3 bucket~~
3. ~~Use AWS Glue to create a database and table with CSV data~~
4. ~~Use Glue ETL to convert the CSV data into following 3~~**~~formats~~** ~~and store the converted files back into a different folder in your bucket.  See important notes below.~~
   1. ~~ORC  (with snappy compression)~~
   2. ~~Parquet   (with snappy compression)~~
   3. ~~Parquet (with gzip compression)~~
5. **~~IMPORTANT NOTE:  For the destination S3 location, don't forget to provide a folder name and trailing slash (/).  If you forget to provide a separate folder name and the ending /, you may get incorrect results and lose points.~~**
6. **~~IMPORTANT NOTE: The use of different compression is not in the recording or in the PDF file.  You will have to look at the Compression Type dropdown in Glue.~~**
7. ~~Verify that the files are created in S3 using the browser. Take screenshot~~
8. ~~Enable the CLI keys~~
9. ~~Using your PC/Mac, run the AWS CLI to get the sizes of the files stored in S3~~
10. ~~Compare the Airline data in S3 stored in all the formats :  CSV, ORC and Parquet formats (including the compressed ones)~~
11. ~~What did you notice about the file sizes, i.e. w~~**~~hich compression (ex. Snappy/gzip) and file format (ORC/Parquet) combination would you recommend for reducing storage costs?~~**

**~~(-5 points if you only write one sentence)~~**

1. ~~Make sure that you have taken all screen shots for this assignment~~
2. ~~Delete the AWS Glue database and tables and S3 buckets used by Glue~~

**Option #2 :  Using Spark to read/convert data ( in Bitnami Hadoop VM in VirtualBox)**

1. Refer to these links for this part of the assignment:
   * <https://spark.apache.org/docs/latest/sql-data-sources-parquet.html#data-source-option>
   * <https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrameWriter.parquet.html>
   * <https://spark.apache.org/docs/3.1.3/api/python/reference/api/pyspark.sql.DataFrameWriter.orc.html>
2. After you load the data into spark, print the schema of the data using printSchema() and then print the count of rows using the count() function call.
3. Use Spark functions to convert the CSV file into the following formats.  (see the slides for the syntax)   
     
   **NOTE:  Use the appropriate function call to convert into ORC/Parquet by looking at the websites given above.  
   NOTE: The use of different compression is not in the recording. You will need to look at the links given above and the slides.   
   NOTE: This assignment is about compression and not about encryption.**
   1. Parquet  (with **gzip** compression)
   2. ORC (with **snappy** compression)
4. Note:  The ORC and Parquet files will be stored in sub-folders.
5. **Important**:  Run the **du -h .** command in the folder that has the CSV, ORC and Parquet data and then compare the file sizes.  They will not be the same and the ORC/Parquet file sizes will be in MB if the original file is more than 20 MB in size.
6. Take screenshots of the file sizes on disk.
   * A screen shot of a computer

     Description automatically generated
7. What did you notice about the file sizes, i.e. w**hich compression (ex. Snappy/gzip) and file format (ORC/Parquet) combination would you recommend for reducing storage costs? (-5 points if you only write one sentence)**
   * **The first thing that I noticed, in general, was that compression significantly reduces the size of a file. But I also noticed that there different methods and effectiveness/efficiency of compression and the choice will affect the file size.**
8. **In this step, you will view the contents of a file and answer the questions.**
   * Download the file called “mysteryfile\_sp242\_004.parq” from the assignment files
   * Transfer the parquet file into bitnami VM using MobaXterm or scp (for Macs)
   * Run the Linux **file** command with the mystery file, which will tell you the type of the file.
   * Start pyspark (preferably in a separate tab window)
   * Read the mystery parquet file using the spark function (see slides for how to read a parquet file)
   * Answer the following questions about the data file. Take screenshots  
     1. How many rows are there in the mystery file?
        1. A black background with white text

           Description automatically generated
     2. What are the columns in the mystery file?
        1. A black screen with white text

           Description automatically generated
     3. What was the output of the file command?
        1. 

1. Make sure that you have taken all screen shots for this assignment
2. **Do not submit the data files (CSV/ORC/Parquet) files.**
3. Stop the VM

**Requirements for the assignments:**

* The assignment write-up must contain a brief description of what you did along with screen shots of the installation and execution of your programs.
* Papers and write-up must be submitted as Word or PDF.
* All screenshots must be included in one document.
* **Providing separate PNG/JPG files will not be accepted**
* **Providing pictures of the screen taken with your mobile phone will not be accepted.**