1. The file is transfer to the instance.

A screenshot of a computer

Description automatically generated

1. Unzip of file.

A computer screen shot of a computer screen

Description automatically generated

1. The file with head.

A screenshot of a computer

Description automatically generated

1. After removing the header from the 2nd and 3rd file.

A screenshot of a computer

Description automatically generated

1. Using cat command.

A screenshot of a computer

Description automatically generated

1. The count of line in the file is listed below.

A screenshot of a computer

Description automatically generated

1. Uploaded the file to s3 bucket.

A screenshot of a computer

Description automatically generated

1. Table is created.

A screenshot of a computer

Description automatically generated

1. CVS to ORC conversion.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. CVS to parquete gzip and parquete snappy.

A screenshot of a computer

Description automatically generated

1. **What did you notice about the file sizes?**

I observed a significant reduction in file sizes after converting the CSV file to different formats. Specifically:

* **CSV**: 57.3 MB
* **ORC**: 1.4 MB
* **Parquet (Snappy Compression)**: 1.7 MB
* **Parquet (Gzip Compression)**: 1.2 MB

This reduction indicates that ORC and Parquet formats, with various compression methods, offer substantial storage efficiency compared to the original CSV format.

**What knowledge did you gain after doing this assignment?**

Through this assignment, I gained insights into the impact of data formats and compression on storage efficiency. Converting data to optimized formats like ORC and Parquet can significantly reduce storage requirements, which is crucial for efficient data management and cost savings in cloud environments.

1. Deleted the database and tables.

A screenshot of a computer

Description automatically generated

|  |  |
| --- | --- |
| Your Name | Sai Kumar Murarishetti |
| Student ID | 30079224 |
|  |  |
| Which option did you use for this assignment? (Glue or Spark) | Glue |
| Month and Year of the data files | Sept2012, October 2012, November 2012 |
| Number of lines in the combined CSV file | 1493460 |
| Size of the combined CSV file | 57.3 |
|  |  |
| For Option #1 (AWS Glue) |  |
| Size of the **Parquet** file with **snappy** compression | 1.7MB |
| Size of the ORC file with **snappy** compression | 1.4MB |
| Size of the **Parquet** file with **gzip** compression | 1.2MB |