

Sai Kumar Murarishetti  
Student ID: 30079224

Student ID: 30079224

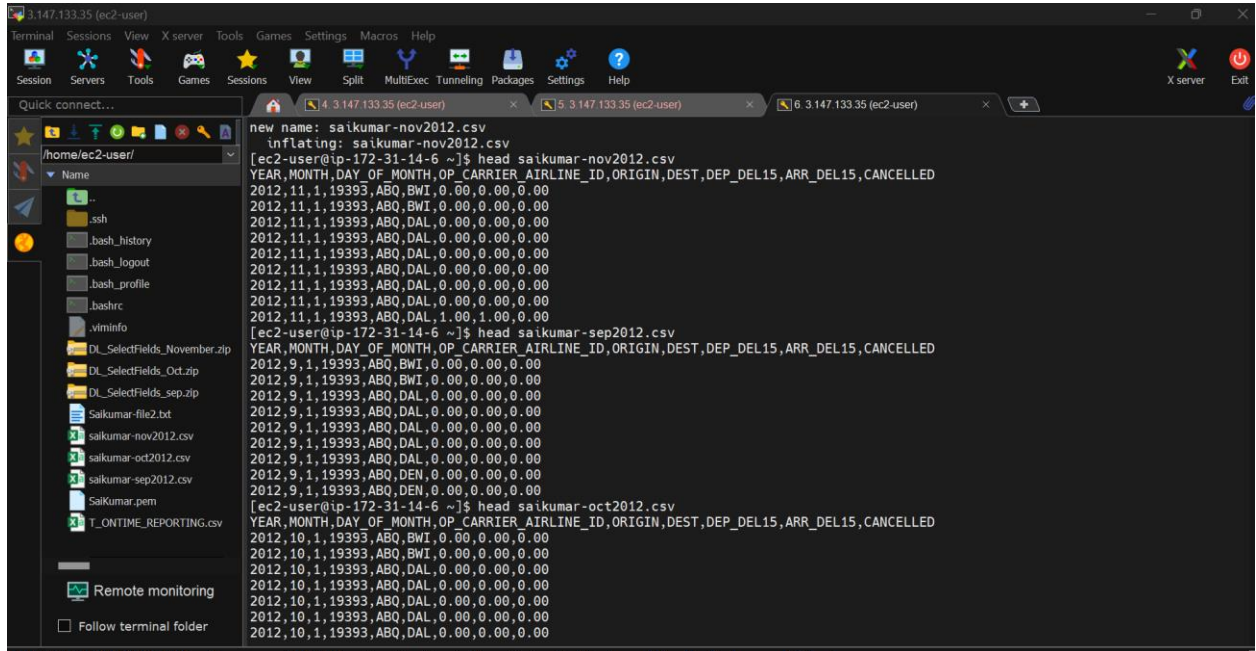
- 
- 3.147.133.35 (ec2-user)
- Terminal Sessions View X server Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
- Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
- Quick connect...
3. 3.147.133.35 (ec2-user)
- MobaXterm Personal Edition v23.5 •  
(SSH client, X server and network tools)
- SSH session to **ec2-user@3.147.133.35**
    - Direct SSH : ✓
    - SSH compression : ✓
    - SSH-browser : ✓
    - X11-forwarding : ✗ (disabled or not supported by server)
  - For more [info](#), ctrl+click on [help](#) or visit our [website](#).
- Amazon Linux 2023
- <https://aws.amazon.com/linux/amazon-linux-2023>
- Last login: Sun Dec 10 06:31:42 2023 from 172.59.216.196  
[ec2-user@ip-172-31-14-6 ~]\$ ls  
DL\_SelectFields\_November.zip DL\_SelectFields\_Oct.zip DL\_SelectFields\_sep.zip Saikumar.pem Saikumar-file2.txt  
[ec2-user@ip-172-31-14-6 ~]\$
- Remote monitoring
- ☐ Follow terminal folder

- 
- 3.147.133.35 (ec2-user)
- Terminal Sessions View X server Tools Games Settings Macros Help
- Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
- Quick connect...
4. 3.147.133.35 (ec2-user) 5. 3.147.133.35 (ec2-user)
- /home/ec2-user/
- Name
- ..
  - .ssh
  - .bash\_history
  - .bash\_logout
  - .bash\_profile
  - .bashrc
  - .viminfo
  - DL\_SelectFields\_November.zip
  - DL\_SelectFields\_Oct.zip
  - DL\_SelectFields\_sep.zip
  - Saikumar-file2.txt
  - saikumar-nov2012.csv
  - saikumar-sep2012.csv
  - Saikumar.pem
  - T\_ONTIME\_REPORTING.csv
- Remote monitoring
- ☐ Follow terminal folder
- ```
[ec2-user@ip-172-31-14-6 ~]$ unzip
.ssh/
[ec2-user@ip-172-31-14-6 ~]$ ls
DL_SelectFields_November.zip  DL_SelectFields_sep.zip  Saikumar-file2.txt  saikumar-nov2012.csv
DL_SelectFields_Oct.zip     Saikumar.pem            T_ONTIME_REPORTING.csv  saikumar-sep2012.csv
[ec2-user@ip-172-31-14-6 ~]$ unzip DL_SelectFields_Oct.zip
Archive:  DL_SelectFields_Oct.zip
replace T_ONTIME_REPORTING.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: r
new name: saikumar-oct2012.csv
  inflating: saikumar-oct2012.csv
[ec2-user@ip-172-31-14-6 ~]$ head saikumar-oct2012.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DEL15,ARR_DEL15,CANCELLED
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
[ec2-user@ip-172-31-14-6 ~]$
```
- UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

## Week 7

Sai Kumar Murarishetti  
Student ID: 30079224

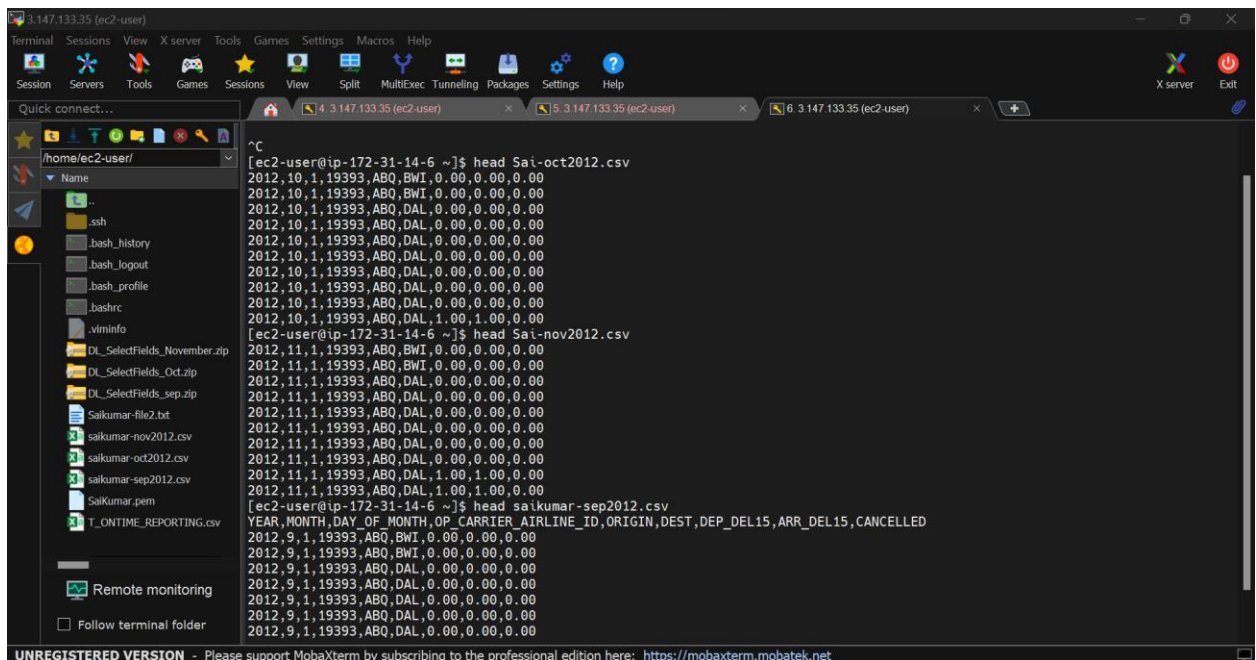
### 3. The file with head.



The screenshot shows a MobaXterm terminal window with three tabs. The active tab is '4. 3.147.133.35 (ec2-user)'. The terminal displays the following commands and output:

```
new name: saikumar-nov2012.csv
inflating: saikumar-nov2012.csv
[ec2-user@ip-172-31-14-6 ~]$ head saikumar-nov2012.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DEL15,ARR_DEL15,CANCELLED
2012,11,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,11,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,1.00,1.00,0.00
[ec2-user@ip-172-31-14-6 ~]$ head saikumar-sep2012.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DEL15,ARR_DEL15,CANCELLED
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DEN,0.00,0.00,0.00
2012,9,1,19393,ABQ,DEN,0.00,0.00,0.00
[ec2-user@ip-172-31-14-6 ~]$ head saikumar-oct2012.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DEL15,ARR_DEL15,CANCELLED
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
```

### 4. After removing the header from the 2<sup>nd</sup> and 3<sup>rd</sup> file.



The screenshot shows a MobaXterm terminal window with three tabs. The active tab is '4. 3.147.133.35 (ec2-user)'. The terminal displays the following commands and output:

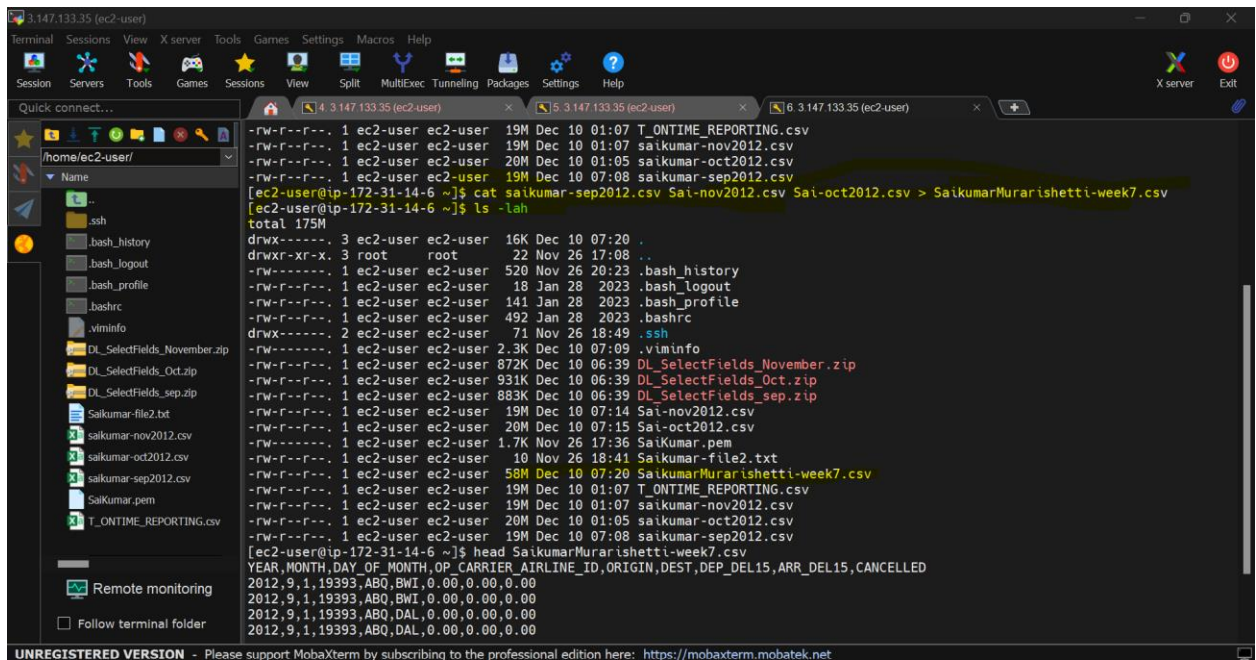
```
^C
[ec2-user@ip-172-31-14-6 ~]$ head Sai-oct2012.csv
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,10,1,19393,ABQ,DAL,1.00,1.00,0.00
[ec2-user@ip-172-31-14-6 ~]$ head Sai-nov2012.csv
2012,11,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,11,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,11,1,19393,ABQ,DAL,1.00,1.00,0.00
2012,11,1,19393,ABQ,DAL,1.00,1.00,0.00
[ec2-user@ip-172-31-14-6 ~]$ head saikumar-sep2012.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DEL15,ARR_DEL15,CANCELLED
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

## Week 7

Sai Kumar Murarishetti  
Student ID: 30079224

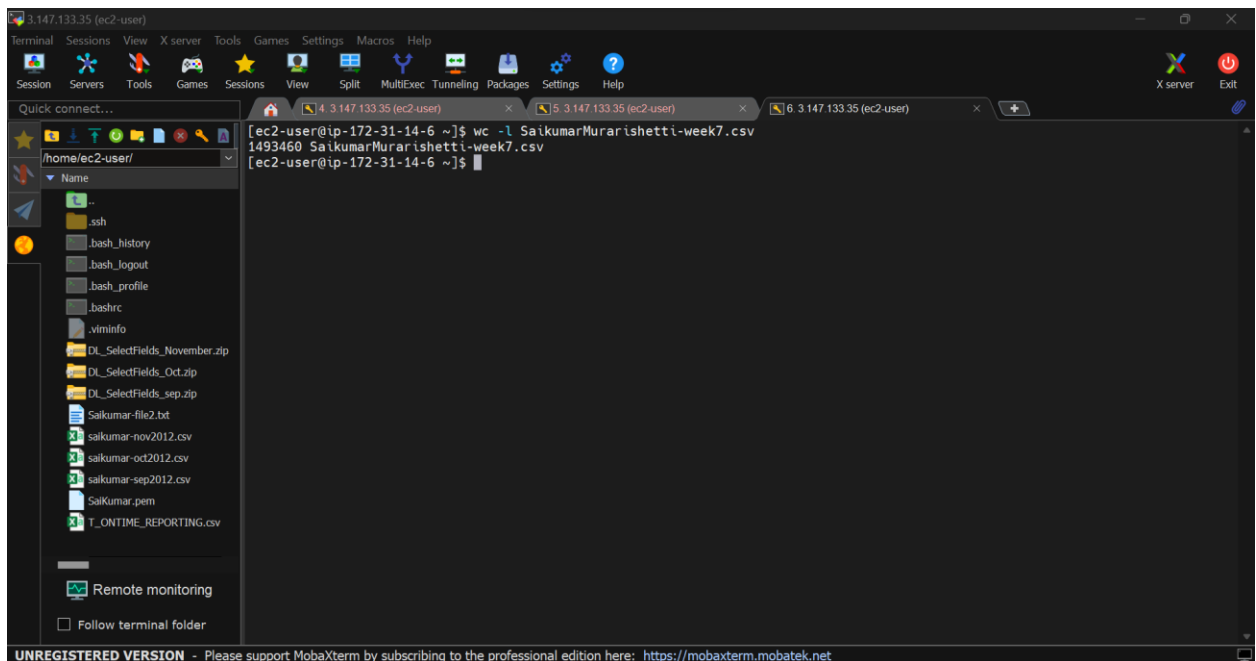
### 5. Using cat command.



The screenshot shows a MobaXterm terminal window with three tabs. The active tab is titled '4. 3.147.133.35 (ec2-user)'. The terminal displays the output of the 'ls -lah' command, listing files and directories in the home directory of the 'ec2-user'. The files include '.bash\_history', '.bash\_logout', '.bash\_profile', '.bashrc', '.viminfo', 'DL\_SelectFields\_November.zip', 'DL\_SelectFields\_Oct.zip', 'DL\_SelectFields\_sep.zip', 'Saikumar-file2.txt', 'saikumar-nov2012.csv', 'saikumar-oct2012.csv', 'saikumar-sep2012.csv', 'Saikumar.pem', and 'T\_ONTIME\_REPORTING.csv'. The output of the 'cat' command is also visible, showing the contents of 'saikumar-sep2012.csv' and 'Sai-nov2012.csv'. The terminal output is as follows:

```
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 01:07 T_ONTIME_REPORTING.csv
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 01:07 saikumar-nov2012.csv
-rw-r--r-- 1 ec2-user ec2-user 20M Dec 10 01:05 saikumar-oct2012.csv
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 07:08 saikumar-sep2012.csv
[ec2-user@ip-172-31-14-6 ~]$ cat saikumar-sep2012.csv Sai-nov2012.csv > SaikumarMurarishetti-week7.csv
[ec2-user@ip-172-31-14-6 ~]$ ls -lah
total 175M
drwx----- 3 ec2-user ec2-user 16K Dec 10 07:20 .
drwxr-xr-x 3 root root 22 Nov 26 17:08 ..
-rw-r--r-- 1 ec2-user ec2-user 520 Nov 26 20:23 .bash_history
-rw-r--r-- 1 ec2-user ec2-user 18 Jan 28 2023 .bash_logout
-rw-r--r-- 1 ec2-user ec2-user 141 Jan 28 2023 .bash_profile
-rw-r--r-- 1 ec2-user ec2-user 492 Jan 28 2023 .bashrc
drwx----- 2 ec2-user ec2-user 71 Nov 26 18:49 .ssh
-rw-r--r-- 1 ec2-user ec2-user 2.3K Dec 10 07:09 .viminfo
-rw-r--r-- 1 ec2-user ec2-user 872K Dec 10 06:39 DL_SelectFields_November.zip
-rw-r--r-- 1 ec2-user ec2-user 931K Dec 10 06:39 DL_SelectFields_Oct.zip
-rw-r--r-- 1 ec2-user ec2-user 883K Dec 10 06:39 DL_SelectFields_sep.zip
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 07:14 Sai-nov2012.csv
-rw-r--r-- 1 ec2-user ec2-user 20M Dec 10 07:15 Sai-oct2012.csv
-rw-r--r-- 1 ec2-user ec2-user 1.7K Nov 26 17:36 Saikumar.pem
-rw-r--r-- 1 ec2-user ec2-user 10 Nov 26 18:41 Saikumar-file2.txt
-rw-r--r-- 1 ec2-user ec2-user 58M Dec 10 07:20 SaikumarMurarishetti-week7.csv
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 01:07 T_ONTIME_REPORTING.csv
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 01:07 saikumar-nov2012.csv
-rw-r--r-- 1 ec2-user ec2-user 20M Dec 10 01:05 saikumar-oct2012.csv
-rw-r--r-- 1 ec2-user ec2-user 19M Dec 10 07:08 saikumar-sep2012.csv
[ec2-user@ip-172-31-14-6 ~]$ head SaikumarMurarishetti-week7.csv
YEAR,MONTH,DAY_OF_MONTH,OP_CARRIER,AIRLINE_ID,ORIGIN,DEST,DEP_DELT5,ARR_DELT5,CANCELLED
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,BWI,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
2012,9,1,19393,ABQ,DAL,0.00,0.00,0.00
```

### 6. The count of line in the file is listed below.



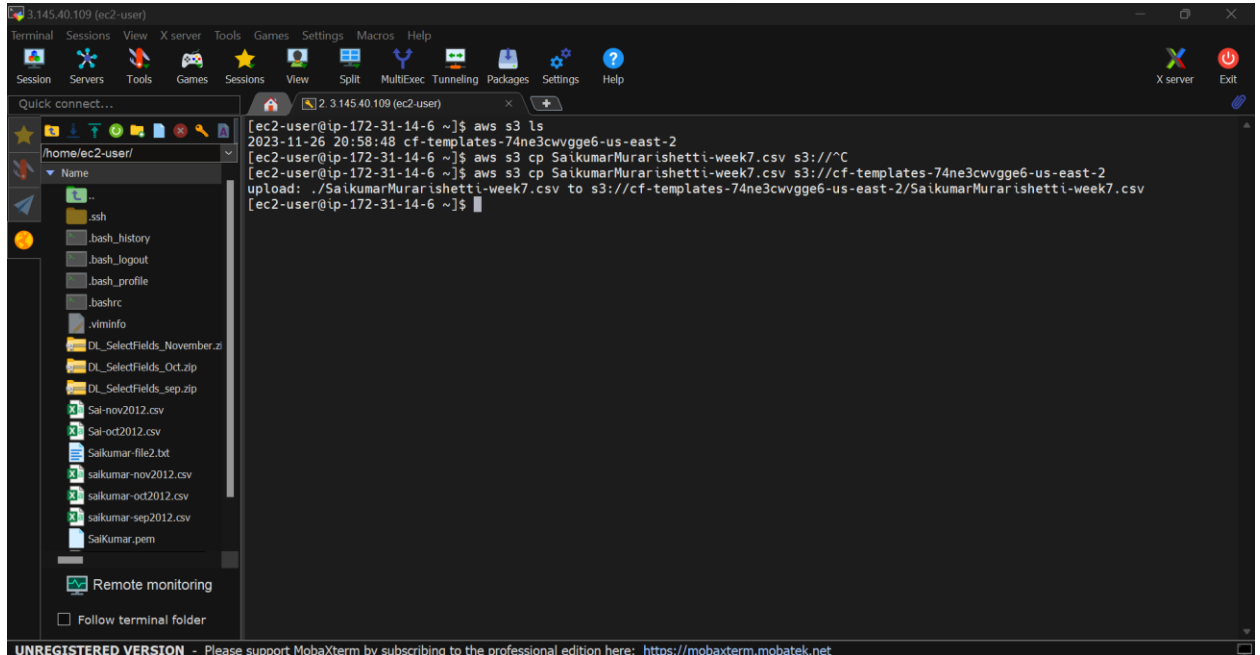
The screenshot shows a MobaXterm terminal window with three tabs. The active tab is titled '4. 3.147.133.35 (ec2-user)'. The terminal displays the output of the 'wc -l' command, which counts the number of lines in the file 'SaikumarMurarishetti-week7.csv'. The output is as follows:

```
[ec2-user@ip-172-31-14-6 ~]$ wc -l SaikumarMurarishetti-week7.csv
1493460 SaikumarMurarishetti-week7.csv
[ec2-user@ip-172-31-14-6 ~]$
```

## Week 7

Sai Kumar Murarishetti  
Student ID: 30079224

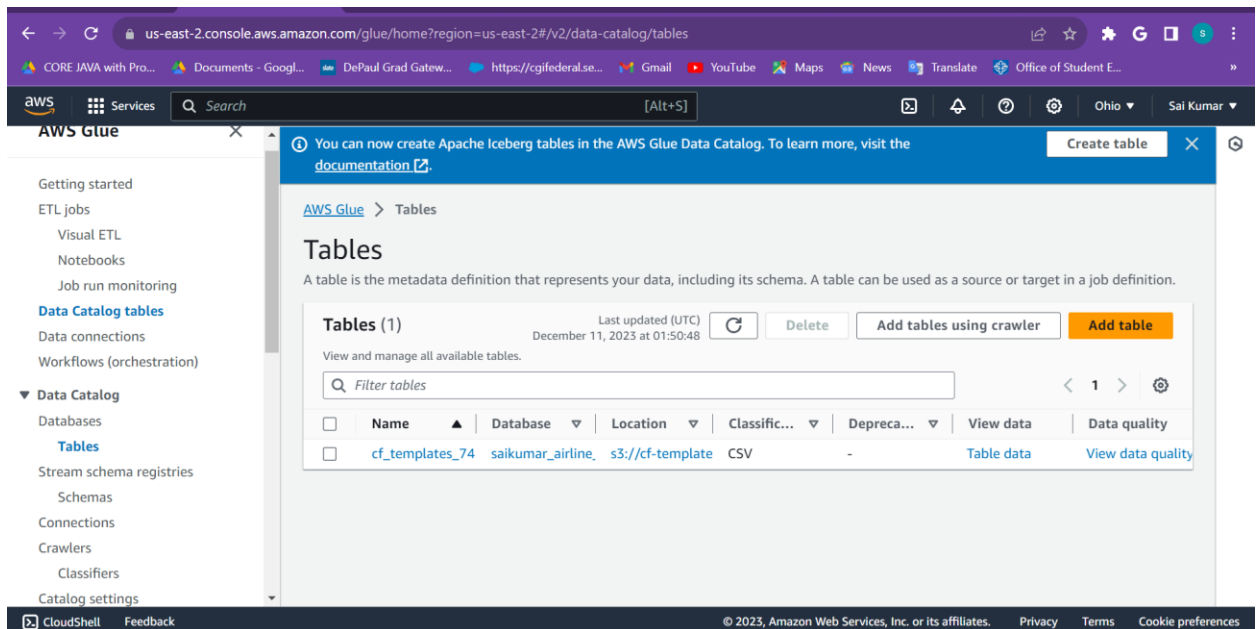
7. Uploaded the file to s3 bucket.



The screenshot shows a MobaXterm terminal window with a file explorer on the left. The terminal displays the following commands and output:

```
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls
2023-11-26 20:58:48 cf-templates-74ne3cwgge6-us-east-2
[ec2-user@ip-172-31-14-6 ~]$ aws s3 cp SaikumarMurarishetti-week7.csv s3://^C
[ec2-user@ip-172-31-14-6 ~]$ aws s3 cp SaikumarMurarishetti-week7.csv s3://cf-templates-74ne3cwgge6-us-east-2
upload: ./SaikumarMurarishetti-week7.csv to s3://cf-templates-74ne3cwgge6-us-east-2/SaikumarMurarishetti-week7.csv
[ec2-user@ip-172-31-14-6 ~]$
```

8. Table is created.



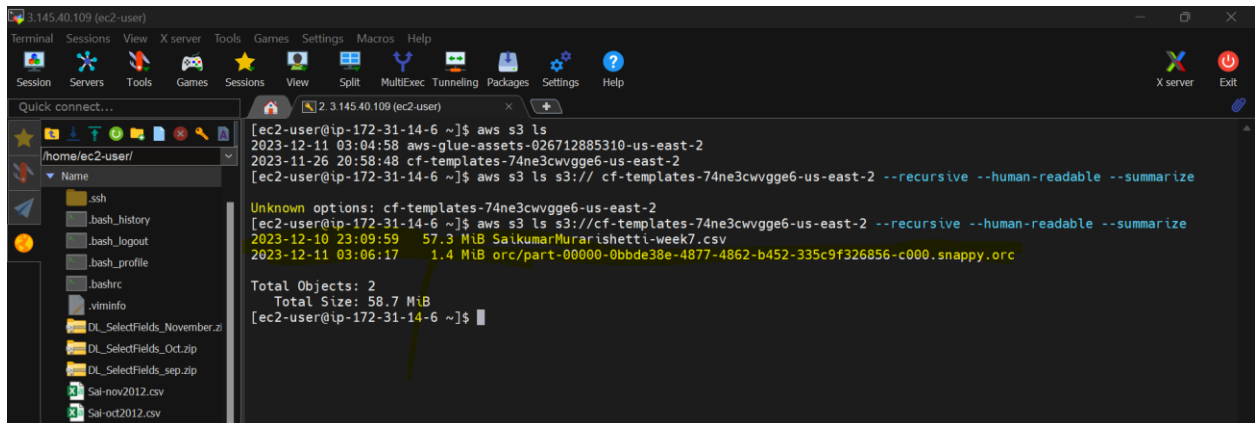
The screenshot shows the AWS Glue console interface. A notification banner at the top states: "You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the [documentation](#)." The left sidebar contains navigation links for Getting started, ETL jobs, Visual ETL, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Data Catalog, Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, and Catalog settings. The main content area displays the "Tables" page, which includes a table listing the available tables.

| Name            | Database         | Location         | Classification | Deprecation | View data                  | Data quality                      |
|-----------------|------------------|------------------|----------------|-------------|----------------------------|-----------------------------------|
| cf_templates_74 | saikumar_airline | s3://cf-template | CSV            | -           | <a href="#">Table data</a> | <a href="#">View data quality</a> |

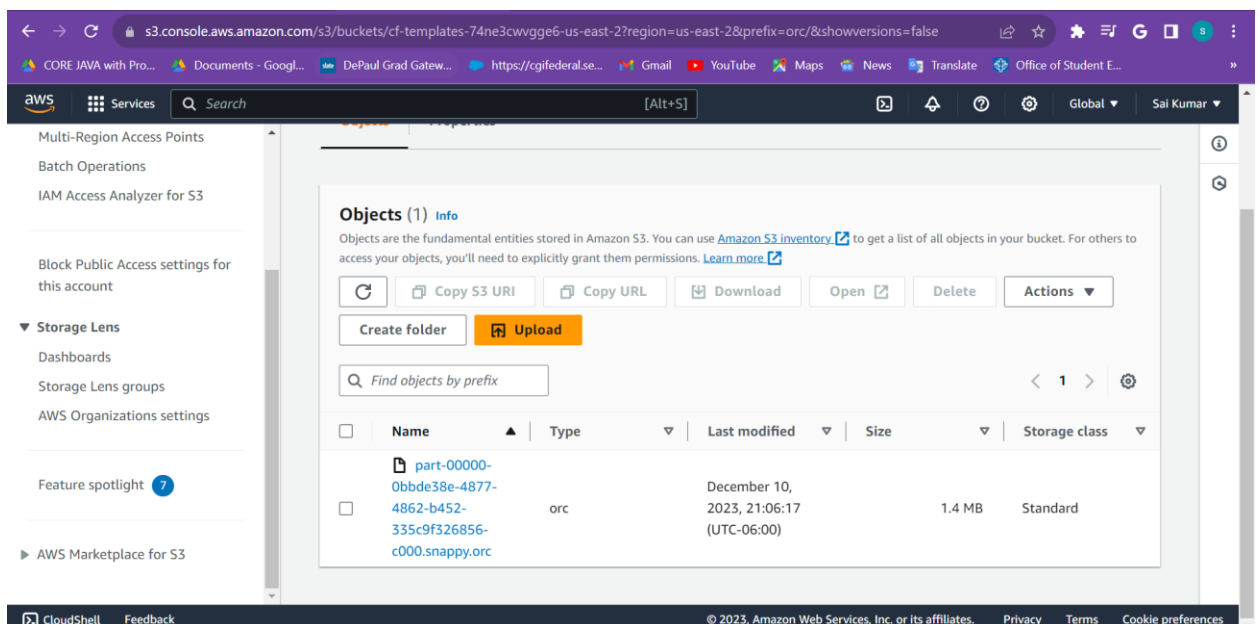
## Week 7

Sai Kumar Murarishetti  
Student ID: 30079224

### 9. CVS to ORC conversion.



```
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls
2023-12-11 03:04:58 aws-glue-assets-026712885310-us-east-2
2023-11-26 20:58:48 cf-templates-74ne3cwgge6-us-east-2
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls s3:// cf-templates-74ne3cwgge6-us-east-2 --recursive --human-readable --summarize
Unknown options: cf-templates-74ne3cwgge6-us-east-2
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls s3://cf-templates-74ne3cwgge6-us-east-2 --recursive --human-readable --summarize
2023-12-10 23:09:59   57.3 MiB SaikumarMurarishetti-week7.csv
2023-12-11 03:06:17   1.4 MiB orc/part-00000-0bbde38e-4877-4862-b452-335c9f326856-c000.snappy.orc
Total Objects: 2
Total Size: 58.7 MiB
[ec2-user@ip-172-31-14-6 ~]$
```



Objects (1) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

| <input type="checkbox"/> | Name                                                                            | Type | Last modified                           | Size   | Storage class |
|--------------------------|---------------------------------------------------------------------------------|------|-----------------------------------------|--------|---------------|
| <input type="checkbox"/> | <a href="#">part-00000-0bbde38e-4877-4862-b452-335c9f326856-c000.snappy.orc</a> | orc  | December 10, 2023, 21:06:17 (UTC-06:00) | 1.4 MB | Standard      |



## 10. CVS to parquet gzip and parquet snappy.

```

[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls
2023-12-11 03:04:38 aws-glue-assets-026712885310-us-east-2
2023-11-26 20:58:48 cf-templates-74ne3cvvgge6-us-east-2
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls s3://cf-templates-74ne3cvvgge6-us-east-2 --recursive --human-readable --summarize
Unknown options: cf-templates-74ne3cvvgge6-us-east-2
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls s3://cf-templates-74ne3cvvgge6-us-east-2 --recursive --human-readable --summarize
2023-12-10 23:09:59    57.3 MiB SaikumarMurarishetti-week7.csv
2023-12-11 03:06:17    1.4 MiB orc/part-00000-0bbde38e-4877-4862-b452-335c9f326856-c000.snappy.orc
Total Objects: 2
Total Size: 58.7 MiB
[ec2-user@ip-172-31-14-6 ~]$ aws s3 ls s3://cf-templates-74ne3cvvgge6-us-east-2 --recursive --human-readable --summarize
2023-12-10 23:09:59    57.3 MiB SaikumarMurarishetti-week7.csv
2023-12-11 03:38:51    1.7 MiB cvstoparquet_snapping/run-1702265776870-part-block-0-r-00000-snappy.parquet
2023-12-11 03:47:48    1.2 MiB cvstoparquetgzip/run-1702266431463-part-block-0-r-00000-gzip.parquet
2023-12-11 03:06:17    1.4 MiB orc/part-00000-0bbde38e-4877-4862-b452-335c9f326856-c000.snappy.orc
Total Objects: 4
Total Size: 61.6 MiB
[ec2-user@ip-172-31-14-6 ~]$

```

## 11. 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.

## Week 7

Sai Kumar Murarishetti  
Student ID: 30079224

12. Deleted the database and tables.

The screenshot shows the AWS Glue console interface. At the top, a green notification banner reads: "One crawler successfully deleted. The following crawler is now deleted: 'saikumardatabasetablecrawler'". Below this, the "Databases" page is displayed, showing "Databases (0)". The page includes a search bar labeled "Filter databases" and a table with columns: Name, Description, Location URI, and Created on (UTC). The table is currently empty, displaying "No resources" and "No resources to display." The left sidebar contains navigation links for "Data Catalog" (Tables, Stream schema registries, Schemas, Connections, Crawlers, Classifiers, Catalog settings) and "Data Integration and ETL" (ETL jobs, Visual ETL, Notebooks, Job run monitoring).

13.

|                                                                |                                       |
|----------------------------------------------------------------|---------------------------------------|
| 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 <b>Parquet</b> file with <b>snappy</b> compression | 1.7MB                                 |
| Size of the ORC file with <b>snappy</b> compression            | 1.4MB                                 |
| Size of the <b>Parquet</b> file with <b>gzip</b> compression   | 1.2MB                                 |