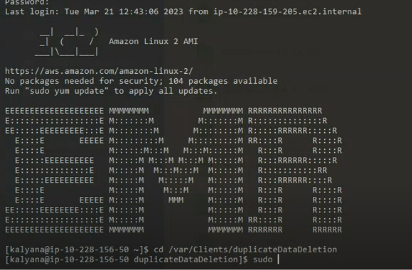
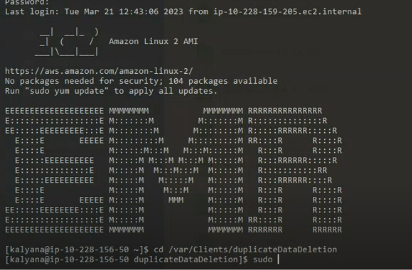
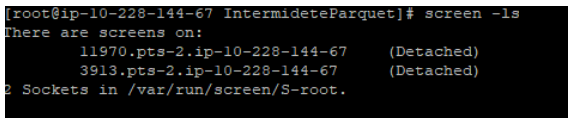
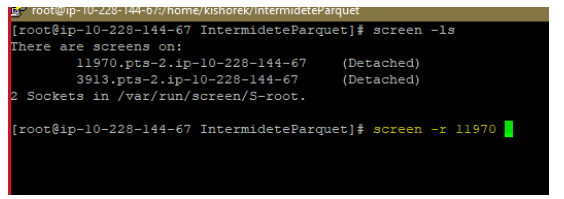
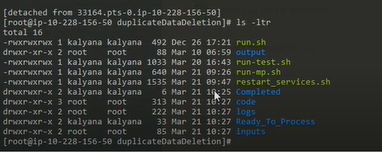
**Hand over of duplicate data deletion**

* We are having duplicate data in osm bucket’s
* Rather then re-creating the whole prqt file again the efficient way is to remove the duplicates and keep only unique data in the parquets
* All the Csv’s we use for the process comes form a sql server.
* It queries each bucket, find how many files are there and shows if any duplicates available and the space it took. Those are imported to CSV and given to us, which we use in emr.
* Box location : [(6) ReadyToProcess | Powered by Box](https://ge.ent.box.com/folder/184053046664)
* These CSV’s have 500 assets days.
* These CSV’s Kalyana will upload to the EMR.
* Need to login into the final-parquet emr in east side AWS account: 564772463473
* **ssh pravin@10.228.156.50**
* 
* And go to /var/Clients/duplicateDataDeletion dir
* 
* **Note:** Start the script after going into screen mode
* Run Screen command (Eg: screen)
* to exit from screen whiteout stopping current executing job we have use ***ctrl+a+d***
* screen can be re-loged in from the same user account where we first started , screen -ls gives the screen numbers open , to login into a particular screen we have to use screen -r number\_of\_the\_screen
* 
* 
* These are the files which are used in the script to delete the data
* ****
* In **Ready\_To\_Process** folder we need to keep those CSV’s
* After we execute the **run-mp.sh** , it goes to input folder and after the process is done the CSV files moved to Completed dir.
* After 1 file is done we have to restart Hadoop node-manager and resource manager , this process cleans up the logs, for this we use **restart\_services.sh** file. It is called inside the main program itself
* The below image is the process flow:
* First the parquet is moved from **osm bucket** to **pd-dev-phhd-output/tmp/tmpDupDataStore**
* 