**Hand over of duplicate data deletion and Array Data**

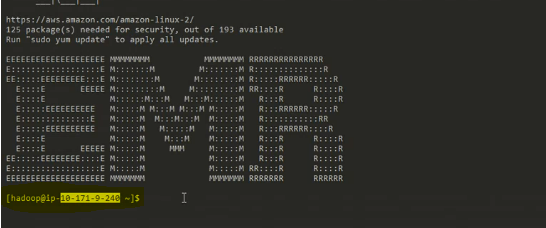
Use case:

**Array tag data request:** As per Analytics team request this task is done, cause Array tags can’t be queried from the APM or other API’s , so that data can be given in csv files

This is raw data array tags, which are given to user when asked

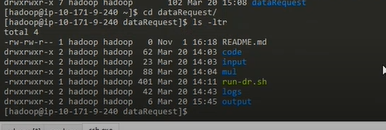
For one data point itself we have multiple Array tags

**How to create these CSV file:** login into the EMR: **usw02-maintenance-pr-ha-blue** master node as shown in below image



**Cmd:** **cd dataRequest/**

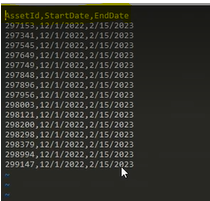
To see the contents of the file: **ls -lrt**



Here we have the script **run-dr.sh** to get the array tag data , before running the script we need to do changes for **inputs** which are used in script , these are in **input** folder.

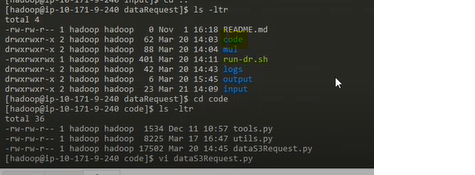


If we go to the **input** folder we have a **input.csv** file, this file has below contents



For number of AssetId’s to get Array data for specific time period we have to give those details as shown in above image and run the script.

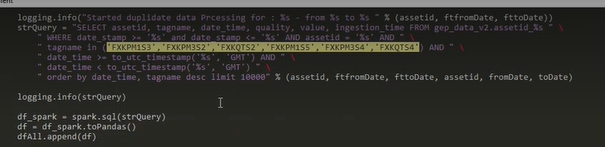
Array tag’s are for now hardcoded if we need to change we need to change we have to do in the code itself and run. Code lies in **code** dir **dataS3Request.py** file as shown below



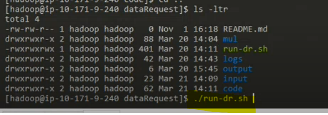
look out for strings or content as shown in below image, For above mentioned file and do the required changes and run

not only array tags even if want to get some other data also we can get from here , cause it is a hive query which we are doing.

**Note:**  This is Raw data only which we are querying here not the interpolated data



Now after doing required changes to get csv file of array tags now we need to run the following shell program **run-dr.sh**

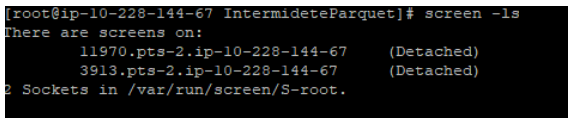


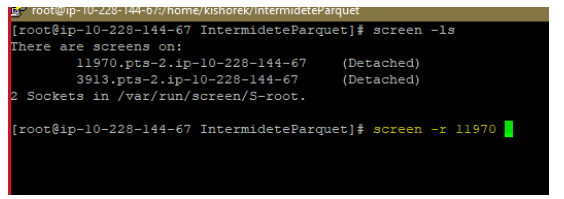
**Note:** Start the script after going into screen mode

Run Screen command (Eg: screen)

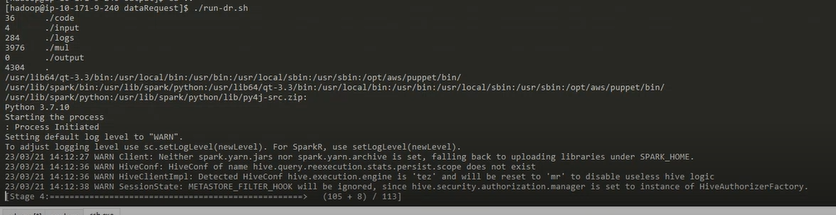
to exit from screen whiteout stopping current executing job we have use ***ctrl+a+d***

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





**./run-dr.sh**



While running this script we get few warnings which can be ignored as the spark version which is in EMR has some limitations

The hive query which runs in background divides the query for 7 day’s and runs , if we see in above screenshot where we give input data. We gave 1 to 2 months data, we cant query it for that long asset days at a time as it will impact the memory usage and any other job’s

The Csv’s are created in the output folder, which is given in below image.

If the csv file size is 1 kb that means there is no data in hive for that requested **array tags-assets**

