CAF sampling guide

This is a guide to take samples from monthly subscribers’ database submitted by TSPs. We

use two Linux codes to achieve automatic sampling and updation of history files of all TSPs at once. It gives 0.1% samples for prepaid and postpaid subscribers and 0.01% for bulk subscribers.

# Prerequisites:

1. Install Cygwin (Linux terminal) from [cygwin 64bit](https://www.cygwin.com/setup-x86_64.exe). ([cygwin home](https://www.cygwin.com/) for 32 bit PC).Choose default options while installing. **See Annexure 1**.
2. Install notepad++ from [here](https://notepad-plus-plus.org/repository/7.x/7.6.3/npp.7.6.3.Installer.x64.exe). Use it to open .sh and .txt files.
3. We need following files

* **folders.sh** – this code quickly creates the required folders to copy the files for sampling.
* **all1.sh** – this code is the main program to sample the databases.
* **paths.txt**- this contains the paths to folders in which sampling is to be done.
* **subscribers’ database** files in **| .txt (pipe delimited .txt)format only**-. Ask TSPs to submit files in this format only otherwise you yourself can export from .mdb(ms- access) to .txt format(see Annexure 2). All rows of the file should not start with any non integer character, though some rows can be there starting with alphabet, blanks or special characters etc and they don’t cause problem but all rows should not start with them . The **first column of file must be mobile numbers**.

Eg. of a line of required file format:-

9424420111|Gjpb447100|Rames Vallabhai |8-8-1963 0:00:00|Vallabhai|4 Manek Baglows Dafnakla Shahibaug|4| ……

Even if number is 9424420111.00 it is not a problem. Decimal part is ignored during sampling by code itself.

Also standardize that first four columns should be those columns which you need to cut from the files.

* **mdn.txt**- this is the history file of numbers that have been sampled. If history file is in any other format other than .txt for eg., history.csv then save this file as mdn.txt **(.txt of MS-DOS**) . It is important that for the very first time use (for a given TSP), the format must be .txt of MS-DOS. For subsequent use, no need to worry about format(see step 6 of Procedure).

# Procedure:

1. **Run Cygwin as** **administrator**.(right click on Cygwin icon in start menu to get the option) Type **pwd** and hit enter to see root directory.
2. Copy all1.sh, folders.sh, paths.txt to that directory. It will be something like this :- C:\cygwin64\home\Sumit
3. Enter the command on Cygwin terminal :- **chmod u+x all1.sh folders.sh**

This command makes the files executable.

1. Steps 2 and 3 are to be done only first time. From subsequent time we can start from this step4 . Enter the command in terminal :- **./folders.sh**
2. Above step creates folders with the name of 5 TSPs in directory:- **C:\Users**

This can be changed by modifying the folders.sh code.(Change this directory from the previous month otherwise it might show error) In the line 2 of this code change directory from **C:\Users** to any desired place. Also you can add or delete a folders for a TSP. Just observe the code for once. Always remember to save .sh(folders.sh) files after editing them.

1. After step 5 go to the TSP folder. For eg. C:\Users\airtel. There you will see 6 folders. First copy the mdn.txt which is the history file in this folder(C:\Users\airtel).Also its good to keep its backup copy with you. . It is important that for the very first time use, the format must be .txt of MS-DOS. To change from .csv to required format Click File>Save As and in the file type choose Text(MS-DOS)(\*.txt). Click save.(If(MS-DOS)(\*.txt) format is not available the you may use Tab delimited DOS .txt format)

(Not for first time user-)If copying mdn.txt by changing the name of mdn\_updated.txt of previous month to mdn.txt of this month, no need to check its .txt format.

1. Copy the database files(should be in pipe “|” delimited .txt format) in the six folders as per requirement. For eg. eKyc files of prepaid subscribers should be copied in:- C:\Users\airtel\pre\_ekyc and likewise. It doesn’t matter if some folders are empty.

**All the files in prepaid and postpaid folders will be sampled 0.1% and all the files in bulk folders will be sampled 0.01%.** Sampling will round off the decimal part in standard way(eg 12.356 to 12, 12.5698 to 13, 12.9863 to 13 and likewise). Suppose you want to take 0.01% and not 0.1% samples of prepaid ekyc files. Then put these files in bulk ekyc folder(it should be empty of any other files) and your result will be in results folder as samples\_bulk\_ekyc.txt with 0.01% sampling of prepaid files.

1. Before moving to next step make sure your directory looks like this:-A screenshot of a cell phone

   Description automatically generated
2. Open paths.txt from C:\cygwin64\home\Sumit in notepad++. Enter the paths of folder TSPs of which you want samples. folders.sh generates folders for 5 TSPs. Enter the paths for folders of TSPs of whom you want to take samples. For eg.

A screenshot of a social media post

Description automatically generated

Paths to directory should start with alphabet only(no space or special charactersor numeric) and paths like **1**:/home/airtel etc should not be given(because it starts with a number). Save the paths.txt file

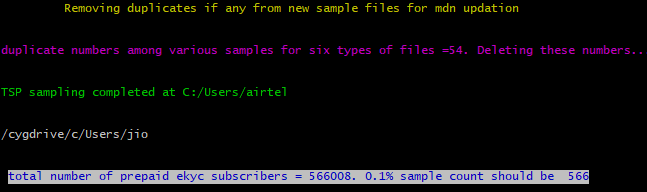
1. Go to Cygwin terminal. Enter the following command:- **./all1.sh |& tee screen.txt** .Hit enter. This will display all the messages on terminal and will also store them in screen.txt in the same directory where you stored all1.sh(C:\cygwin64\home\Sumit). Use notepad++ to view this file. This file can be used for troubleshooting.

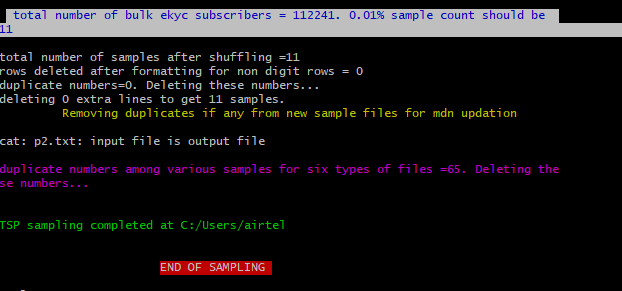
Alternatively, you can also run **./all1.sh** only. This is simpler but you won’t be able to store the terminal screen output.

Also, to check the time taken to run the code you can put ‘time’ before the command. For eg **time** **./all1.sh |& tee screen.txt** . Or simply **time ./all1.sh**

1. Upon completion of sampling process your terminal screen might look like this.







1. You can safely ignore all **cat** error messages as shown below

“cat: '\*.txt': No such file or directory

cat: p2.txt: input file is output file

”

Wherever deleting negative lines message is there, after that we again sample with higher shuffle count till deleting positive value of lines is shown.

1. See your results in folder**C:\Users\airtel\results** for airtel and likewise for other TSPs.
2. If you want sampled file of postpaid non\_ekyc look for file samples\_post\_paper.txt. Some files have size 0 KB. This only means that there we no corresponding files to be sampled in some of the 6 folders.This is ok. mdn\_updated.txt is the new updated mdn file. Its size will be almost similar to initial mdn.txt file. This because if 1000 new numbers are sampled this month then these 1000 numbers are put at top of mdn.txt and 1000 numbers are deleted from the bottom of it. If you don’t want to delete 1000 numbers from bottom of it then put **#** at start of line 304(#mdn\_constant) in all1.sh file and remove # from the start of next(305) line(mdn\_incremental).

save the all1.sh file

. **This file(mdn\_updated.txt) should be used as mdn.txt after changing name mdn\_updated.txt to mdn.txt for sampling in the next month**. See the screen.txt file generated in directory where you stored all1.sh file(C:\cygwin64\home\Sumit) for details of all the process that happened on terminal. It can be used for troubleshooting and cross verification of files stored in results folder.

List of new numbers sampled will be in numbers folder.

1. **This completes our sampling process**.
2. In columns\_selector function in all1.sh file at line 158 set the values of a,b,c,d as the column numbers you want to delete. Put # at line 303 if you don’t know which columns to select. You may cut the columns later after completing sampling, from the sampled files manually. Also,the portion of code - awk -F'[|,;\t]' in lines 159 to 169 assumes the delimiter is pipe/ comma/ semicolon/tab. This should later be changed to | (pipe) only(by replacing [|,;\t] with [|] in each line from 159 to 169. Press Save. ) when TSPs start complying with the format by giving subscriber databse in pipe delimited format only.
3. Unless and until you get this error message “ERROR HAPPENED. First column is not mobile numbers.Ask TSP to submit database files with first column as mobile numbers

and then run the code afresh”. Or your terminal runs into infinite loop or you don’t get the desired sampled size files in result folder, everything is fine . But **if any error happens** or you need to stop the terminal **to redo the sampling** then you should firstly **close the Cygwin terminal**. Then **delete the two folders- results and numbers**. Restart the Cygwin terminal as administrator and enter the command in step 10 ( ./all1.sh |& tee screen.txt ).

1. Also delete the folders created through **folders.sh** so that next month when you run the code, you don’t get the error message that “folder already exists”.(Unless you don’t wish to change the directory for next month sampling).

# Annexure 1: How to install Cygwin?

1. go to [here](https://cygwin.com/install.html) . (https://cy gwin.com/install.html.)A screenshot of a social media post

   Description automatically generated
2. Download setup-x86\_64.exe for 64 bit version .
3. Execute the downloaded setup.

A screenshot of a cell phone

Description automatically generated

1. Click Next

A screenshot of a cell phone

Description automatically generated

1. Click next.

A screenshot of a cell phone

Description automatically generated

1. Click next.

A screenshot of a cell phone

Description automatically generated

1. Click next.

A screenshot of a cell phone

Description automatically generated

1. Click next.

A screenshot of a cell phone

Description automatically generated

1. Click next.

A screenshot of a social media post

Description automatically generated

1. Click next. (No need to install anything in particular. Let all be default options)

A screenshot of a cell phone

Description automatically generated

1. Click Next.

A screenshot of a cell phone

Description automatically generated

1. Now the required packages will be downloaded and automatically installed.
2. Click Finish to complete the installation.
3. Run it as administrator once and then close it.(Some files are to be generated on first time use.)

# Annexure 2: Export .mdb or .accdb (Ms Access) files to | delimted .txt files

1. Go to **Data/External Data** tab.

A screenshot of a social media post

Description automatically generated

1. Click Text file under Export group of icons.

A screenshot of a social media post

Description automatically generated

1. Click OK.

A screenshot of a cell phone

Description automatically generated

1. Note that Delimited option is checked. Click Next.

A screenshot of a social media post

Description automatically generated

1. **Default option would be comma. Change** it to **Other**  and in the box **enter |** (pipe symbol). Click Next.

A screenshot of a cell phone

Description automatically generated

1. It will ask file location to save. Leave it unchanged or you can change it to desired location. Click Finish.
2. **THE EXPORT PROCESS IS COMPLETE HERE.(You may click cancel on next prompt and ignore next steps)**

A screenshot of a cell phone

Description automatically generated

1. Tick ‘save export steps’ . In the Save as and Description boxes you can write as shown above in picture or anything you find convenient. This step is useful only if table name is same in all the files to export. Abort the operation if you don’t understand!! Click cancel!!
2. Next time when you open any database file for conversion . Simply Go to  **step 1** above and select **Saved Exports**.

A screenshot of a social media post

Description automatically generated

1. Click the C:\Users\Sumit\Documents\Table1.txt (whatever path it shows )in the pink box.

Change Table1.txt to mylfilename.txt (your file name for this new .txt file).

This file name must be changed every time you convert a .mdb to .txt file.

A screenshot of a cell phone

Description automatically generated

1. Click run.