**DIAL D4D Design Document**

**Version – 0.1**

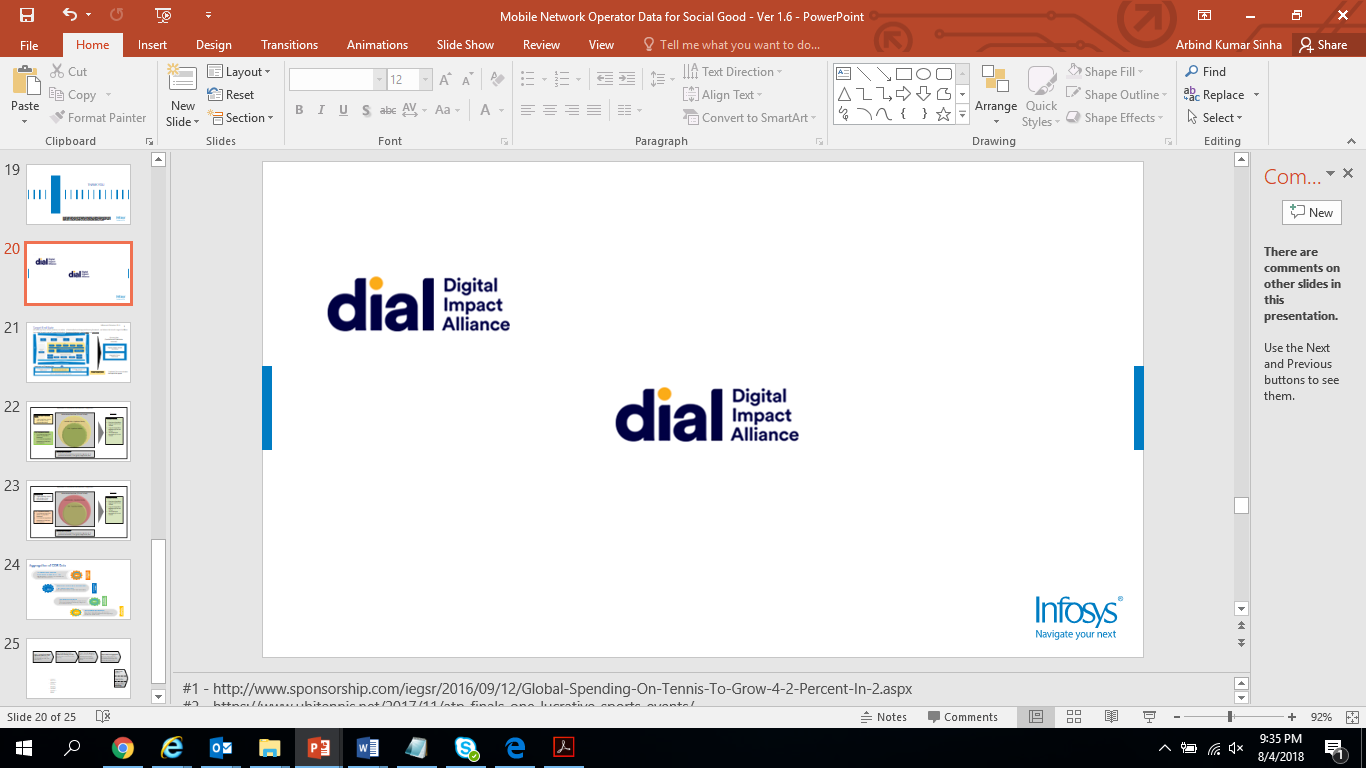


Table of Content

[**DIAL D4D Design Document** 1](#_Toc530666532)

[Introduction 3](#_Toc530666533)

[High Level Data Flow 3](#_Toc530666534)

[Data Flow Architecture 4](#_Toc530666535)

[Data Ingestion 4](#_Toc530666536)

[Malawi server to Azure BLOB 4](#_Toc530666537)

[Azure BLOB to ADL 5](#_Toc530666538)

[Azure ADL to Hive 5](#_Toc530666539)

[Hive Database 8](#_Toc530666540)

[Hive Tables 8](#_Toc530666541)

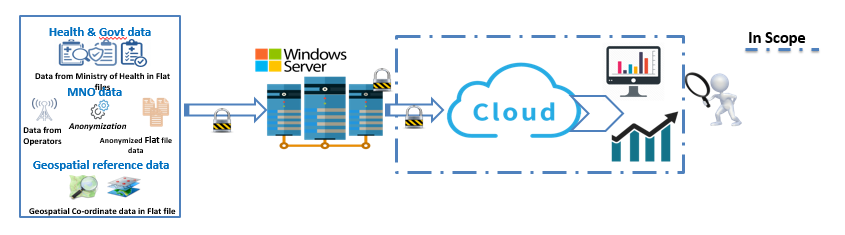
[Reference Documents 13](#_Toc530666542)

# Introduction

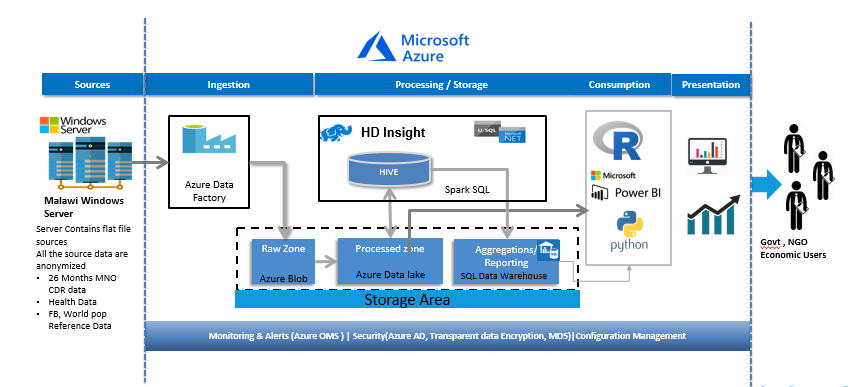
This document covers details of the data model required for Malawi use case execution. This document will also provide details of the tables required and it’s loading logics

## High Level Data Flow

* Data from various sources will be transferred to a central repository in Malawi.
* From the central repository data will be then copied to identified Cloud Services Azure.
* Data processing and analytics will be performed on Azure Cloud
* Only the results will be published to the outside world.



## Data Flow Architecture



# Data Ingestion

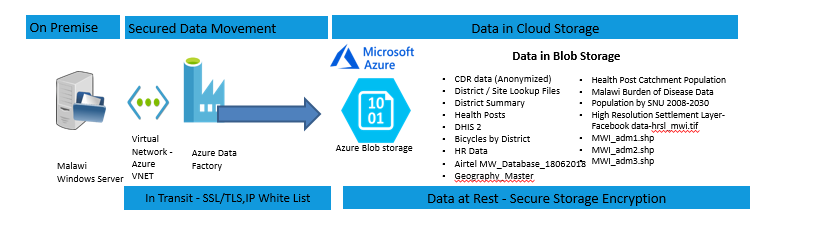
The document should cover from Ingestion (Malawi server to Blob to ADL to Hive)

## Malawi server to Azure BLOB

The Source Data from Malawi Windows server will be transferred to Azure Blob storage.

Below is the two-step process involved.

* 1. Secured Virtual Network will be established to connect the cloud.
  2. Azure data factory will be used as a data pipeline to fetch the data from Malawi server to Azure Blob storage.



## Azure BLOB to ADL

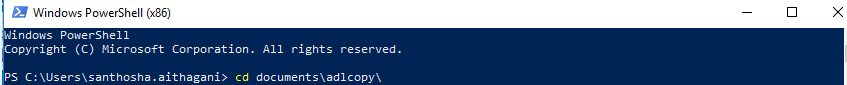
1. Install adlcopy in your local machine.
2. In Powershell go the adlcopy location and execute below command.

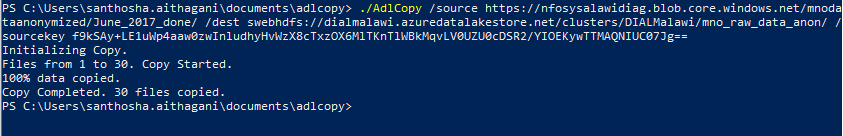
Example:

./AdlCopy /source <https://nfosysalawidiag.blob.core.windows.net/mnodataanonymized/June_2017_done/> /dest swebhdfs://dialmalawi.azuredatalakestore.net/clusters/DIALMalawi/**mno\_raw\_data\_anon**/ /sourcekey f9kSAy+LE1uWp4aaw0zwInludhyHvWzX8cTxzOX6MlTKnTlWBkMqvLV0UZU0cDSR2/YIOEKywTTMAQNIUC07Jg==

Green highlighted – Source file or folder name

Yellow highlighter – Target ADL directory.





## Azure ADL to Hive

Create an external table pointing to the ADL location.

Example:

CREATE EXTERNAL TABLE IF NOT EXISTS **mno\_raw\_data\_anon**(

    USAGE\_TYPE\_NAME STRING

    ,CALL\_ORIGINATING\_NUMBER STRING

    ,CALL\_TERMINATING\_NUMBER STRING

    ,CALL\_START\_DATE STRING

    ,CALL\_START\_TIME STRING

    ,Duration String

    ,LAC\_ID STRING

    ,CELL\_ID STRING)

    COMMENT 'Test data'

    ROW FORMAT DELIMITED

    FIELDS TERMINATED BY ','

    STORED AS TEXTFILE

    location 'adl://dialmalawi.azuredatalakestore.net/clusters/DIALMalawi/**mno\_raw\_data\_anon** ';

**Identify Active subscribers:**

Identify active and inactive numbers from the aggregated set.

* Inactive number: - If the number is not used for 3 months consecutively, mark that number as inactive.
* From home\_location\_monthly table, flag the number which are identified as inactive.
* Script Location: /home/dialssh/dev/active\_inactive\_dev/active\_users.sql

**Identify home\_location:**

Home\_location\_day and home\_location\_month tables have multilple cell\_id’s mapped to single originating number for a particular day and month respectively.

* Load the home\_location\_day table with most used cell\_id for a particular day  and the latest cell\_id used when there is tie in count.
* Load the home\_location\_month table likewise from home\_location\_day.

The no\_of\_unique\_id column in mno\_daily\_agg should be calculated from home\_location\_day table. Likewise for mno\_monthly\_agg from home\_location\_monthhly.

* Script Location: /home/dialssh/dev/agg\_on\_MNO\_FULL\_DATA/10\_homelocation\_day.sql

**Filter for 7pm - 7am records:**

Filtered table with data from 7 pm to 7 am and do aggregation on top of this table.

* Script Location for filtering night records: /home/dialssh/dev/mno\_NIGHT/mno\_cdr\_active\_7pm\_to\_7am.sql
* Script location for aggregation on top of night records: /home/dialssh/dev/agg\_on\_MNO\_FULL\_DATA

**Prepare source tables:**

**Morning hours**: Filter table with data from 7am to 7 pm

* Script Location: /home/dialssh/dev/mno\_MORNING/mno\_7am\_7pm.sql

**Rainy season:** Filter table with data from November to April

* Script Location: /home/dialssh/dev/MNO\_NIGHT\_seasons/MNO\_NIGHT\_RAINY\_SEASON.sql

**non-rainy:** Filter table with data from May to October

* Script Location: /home/dialssh/dev/MNO\_NIGHT\_seasons/MNO\_NIGHT\_NON\_RAINY\_SEASON.sql

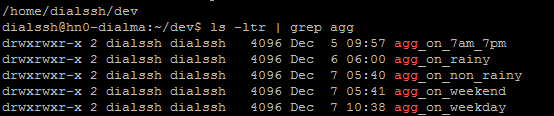
**weekend:** Filter table for weekends

* Script Location: /home/dialssh/dev/MNO\_NIGHT\_seasons/MNO\_NIGHT\_WEEKEND.sql

**weekday:** Filter table for weekdays

* Script Location: /home/dialssh/dev/MNO\_NIGHT\_seasons/MNO\_NIGHT\_WEEKDAY.sql

**Aggregation Rules on top of 5 source tables:** Apply aggregation on all the source tables

* Script Location: /home/dialssh/dev 

**Long term population movement:**

*Step1 :* Create a new table **homelocation\_monthly\_TA** by mapping existing 2 tables **homelocation\_monthly** and **cell\_id\_withoutlocation.**

**Script Location:** /home/dialssh/dev/long\_term\_pop\_scripts/homelocation\_tables/homelocation\_monthly\_TA.sql

*Step2:-* Create 2 tables by transforming the above table. Each user to have 1 record with district mapping for each month in one table and TA mapping for each month in another table.

**Script Location:**

/home/dialssh/dev/long\_term\_pop\_scripts/homelocation\_tables/user\_homelocation\_monthly\_TA.sql

/home/dialssh/dev/long\_term\_pop\_scripts/homelocation\_tables/user\_homelocation\_monthly\_district.sql

**Short term population movement:**

This is to calculate the movement of the people for very short term over a week during the different seasons like rainy, non-rainy and weekend on the day data between 7AM to 9PM for every 2 hours as per the below table.

**Script Location:** /home/dialssh/dev/short\_term\_pop\_scripts

**Steps to calculate the short term population:**

* Get the active user whose is active more than 4 hours per day records from the Main CDR table

Script Location:

/home/dialssh/dev/short\_term\_pop\_scripts/ active\_4\_hr\_bulk.sql

/home/dialssh/dev/short\_term\_pop\_scripts/ active\_4\_hr\_filtered.sql

/home/dialssh/dev/short\_term\_pop\_scripts/ active\_4\_hr\_filtered\_part.sql

* Identify the data for short\_term\_pop\_14days:

Script Location: /home/dialssh/dev/short\_term\_pop\_scripts/short\_term\_pop\_14days.sql

* short\_term\_pop\_14days\_Timeslot: Partition the table into consecutive 2 hour timeslots

Script Location:

/home/dialssh/dev/short\_term\_pop\_scripts/ short\_term\_pop\_14days\_Timeslot.sql

* short\_term\_lvl1

Script Location: /home/dialssh/dev/short\_term\_pop\_scripts/short\_term\_lvl1.sql

* short\_term\_Master

Script Location: /home/dialssh/dev/short\_term\_pop\_scripts/short\_term\_Master.sql

**Rainy season:**

Table Name: short\_term\_pop\_movement\_Rainy\_season

Script Location: /home/dialssh/dev/short\_term\_pop\_scripts/short\_term\_pop\_movement\_Rainy\_season.sql

Column Names:

call\_originating\_number string   
7amto9am string  
9amto11am string   
11amto13pm string   
13pmto15pm string   
15pmto17pm string   
17pmto19pm string   
19pmto21pm string   
call\_start\_date date

**Non-Rainy season:**

Table Name: short\_term\_pop\_movement\_Non\_Rainy\_season

Script Location:

/home/dialssh/dev/short\_term\_pop\_scripts/ short\_term\_pop\_movement\_Non\_Rainy\_season.sql

Column Names:

call\_originating\_number string   
7amto9am string  
9amto11am string   
11amto13pm string   
13pmto15pm string   
15pmto17pm string   
17pmto19pm string   
19pmto21pm string   
call\_start\_date date

**Weekends:**

Table Name: short\_term\_pop\_movement\_Weekend

Script Location:

/home/dialssh/dev/short\_term\_pop\_scripts/ short\_term\_pop\_movement\_Weekend.sql

Column Names:

call\_originating\_number string   
7amto9am string  
9amto11am string   
11amto13pm string   
13pmto15pm string   
15pmto17pm string   
17pmto19pm string   
19pmto21pm string   
call\_start\_date date

# Hive Database

## Hive Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SL No** | **database** | **Table name** | **Field names** | **Description** | **Loading Logic** |
| 1 | etldb | mno\_cdr\_data\_cleansed | usage\_type\_name string, call\_originating\_number string, call\_terminating\_number string, call\_start\_date string, call\_start\_time string , duration string, lac\_id string, cell\_id string | Contains all the CDRs | 1:1 mapping from source |
| 2 | etldb | active\_cdr\_cleansed | usage\_type\_name string  call\_originating\_number string  call\_terminating\_number string  call\_start\_date date  call\_start\_time timestamp  duration string  lac\_id string  cell\_id string | Contains active records with call\_start\_time in timestamp format | The call originating numbers which appears on last 3 months are considered as Active Numbers.  Those numbers will be joined with complete MNO data to fetch CDR for the active numbers |
| 3 | etldb | mno\_cdr\_active\_7pm\_to\_7am | usage\_type\_name string  call\_originating\_number string  call\_terminating\_number string  call\_start\_date date  call\_start\_time timestamp  duration string  lac\_id string  cell\_id string | Active records which are time filtered ( 7PM - 7 AM) | Filtered call\_start\_time column to have only records from 7pm to 7am |
| 4 | etldb | AGG\_DAY\_WISE | CALL\_ORIGINATING\_NUMBER string  CALL\_START\_DATE date  CELL\_ID string  NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level | Day wise cell\_id level aggregation done on mno\_cdr\_active\_7pm\_to\_7am |
| 5 | etldb | AGG\_MONTH\_WISE | CALL\_ORIGINATING\_NUMBER string  CALL\_MONTH\_YEAR string  CELL\_ID string  NO\_OF\_EVENTS int | Month wise aggregation on each subscriber at cell\_id level | Month wise cell\_id level aggregation done on mno\_cdr\_active\_7pm\_to\_7am |
| 6 | etldb | HOMELOCATION\_DAY | CALL\_ORIGINATING\_NUMBER string  CALL\_START\_DATE date  CALL\_START\_TIME timestamp  CELL\_ID string  RANK int | Home location for each number | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 7 | etldb | HOMELOCATION\_MONTHLY | CALL\_ORIGINATING\_NUMBER STRING ,CALL\_START\_DATE DATE ,MONTH\_YEAR STRING ,CELL\_ID STRING ,RANK INT | Month wise homelocation for each active subscriber | For each number in a given month , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 8 | etldb | MNO\_DAILY\_AGG\_cdr | CELL\_ID STRING ,CALL\_START\_DATE STRING ,NO\_OF\_EVENTS INT | Day wise count of events at each tower | This table calculates the no of subscribers calling from a particulat tower in a single day. |
| 9 | etldb | MNO\_DAILY\_AGG\_HOME | CELL\_ID STRING ,CALL\_START\_DATE STRING ,NO\_OF\_UNIQUE\_ID INT | Distinct day wise records for each cell\_id | This table calculates the distinct records per day at cell\_id level from homelocation table. |
| 10 | etldb | MNO\_DAILY\_AGG | CELL\_ID STRING ,CALL\_START\_DATE STRING ,NO\_OF\_EVENTS INT ,NO\_OF\_UNIQUE\_ID INT | Day wise aggregstion on cell\_id level | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |
| 11 | etldb | MNO\_MONTHLY\_HOME | CELL\_ID STRING  ,MONTH\_YEAR STRING  ,NO\_OF\_UNIQUE\_ID INT | Distinct month wise records for each cell\_id | Distict records calculated from homelocation table on cell\_id level |
| 12 | etldb | MNO\_MONTHLY\_CDR | CELL\_ID STRING  ,MONTH\_YEAR STRING  ,NO\_OF\_EVENTS INT | Month wise count of events at cell\_id level | This table calculates the no of subscribers calling from a particulat tower monthly. |
| 13 | etldb | MNO\_MONTHLY\_AGG | CELL\_ID STRING  ,CALL\_MONTH\_YEAR STRING  ,NO\_OF\_EVENTS INT  ,NO\_OF\_UNIQUE\_ID INT | Month wise aggregation on cell\_id level with unique records from homelocation table | Month wise aggregation on cell\_id level with unique records from homelocation table |
| 14 | etldb | short\_term\_pop\_movement\_Rainy\_season | call\_originating\_number string  ,7amto9am string  ,9amto11am string  ,11amto13pm string  ,13pmto15pm string  ,15pmto17pm string  ,17pmto19pm string  ,19pmto21pm string  ,call\_start\_date date | short term population movement for one week of day data in rainy season on every two hours | short term population movement for one week of day data in rainy season on every two hours |
| 15 | etldb | short\_term\_pop\_movement\_Non\_Rainy\_season | call\_originating\_number string  ,7amto9am string  ,9amto11am string  ,11amto13pm string  ,13pmto15pm string  ,15pmto17pm string  ,17pmto19pm string  ,19pmto21pm string  ,call\_start\_date date | short term population movement for one week of day data in Non-rainy season on every two hours | short term population movement for one week of day data in Non-rainy season on every two hours |
| 16 | etldb | short\_term\_pop\_movement\_Weekend | call\_originating\_number string  ,7amto9am string  ,9amto11am string  ,11amto13pm string  ,13pmto15pm string  ,15pmto17pm string  ,17pmto19pm string  ,19pmto21pm string  ,call\_start\_date date | short term population movement for two week end of day data in rainy and non-rainy season on every two hours | short term population movement for two week end of day data in rainy and non-rainy season on every two hours |
| 17 | etldb | AGG\_DAY\_WISE\_rainy | CALL\_ORIGINATING\_NUMBER string ,CALL\_START\_DATE date  ,CELL\_ID string ,NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level on rainy season | Day wise cell\_id level aggregation done on MNO\_NIGHT\_RAINY\_SEASON |
| 18 | etldb | HOMELOCATION\_DAY\_rainy | CALL\_ORIGINATING\_NUMBER string  ,CALL\_START\_DATE date  ,CALL\_START\_TIME timestamp  ,CELL\_ID string  ,RANK int | Home location for each number on rainy season | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 19 | etldb | MNO\_DAILY\_AGG\_rainy | CELL\_ID string  ,CALL\_START\_DATE string  ,NO\_OF\_EVENTS int  ,NO\_OF\_UNIQUE\_ID int | Day wise aggregstion on cell\_id level on rainy season | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |
| 20 | etldb | AGG\_DAY\_WISE\_non\_rainy | CALL\_ORIGINATING\_NUMBER string ,CALL\_START\_DATE date  ,CELL\_ID string ,NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level on non-rainy season | Day wise cell\_id level aggregation done on MNO\_NIGHT\_NON\_RAINY\_SEASON |
| 21 | etldb | HOMELOCATION\_DAY\_non\_rainy | CALL\_ORIGINATING\_NUMBER string  ,CALL\_START\_DATE date  ,CALL\_START\_TIME timestamp  ,CELL\_ID string  ,RANK int | Home location for each number on non-rainy season | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 22 | etldb | MNO\_DAILY\_AGG\_non\_rainy | CELL\_ID string  ,CALL\_START\_DATE string  ,NO\_OF\_EVENTS int  ,NO\_OF\_UNIQUE\_ID int | Day wise aggregstion on cell\_id level on non-rainy season | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |
| 23 | etldb | AGG\_DAY\_WISE\_weekend | CALL\_ORIGINATING\_NUMBER string ,CALL\_START\_DATE date  ,CELL\_ID string ,NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level on 6week end | Day wise cell\_id level aggregation done on MNO\_NIGHT\_WEEKEND |
| 24 | etldb | HOMELOCATION\_DAY\_weekend | CALL\_ORIGINATING\_NUMBER string  ,CALL\_START\_DATE date  ,CALL\_START\_TIME timestamp  ,CELL\_ID string  ,RANK int | Home location for each number on week end | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 25 | etldb | MNO\_DAILY\_AGG\_weekend | CELL\_ID string  ,CALL\_START\_DATE string  ,NO\_OF\_EVENTS int  ,NO\_OF\_UNIQUE\_ID int | Day wise aggregstion on cell\_id level on week end | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |
| 26 | etldb | AGG\_DAY\_WISE\_weekday | CALL\_ORIGINATING\_NUMBER string ,CALL\_START\_DATE date  ,CELL\_ID string ,NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level on week day | Day wise cell\_id level aggregation done on MNO\_NIGHT\_WEEKDAY |
| 27 | etldb | HOMELOCATION\_DAY\_weekday | CALL\_ORIGINATING\_NUMBER string  ,CALL\_START\_DATE date  ,CALL\_START\_TIME timestamp  ,CELL\_ID string  ,RANK int | Home location for each number on week end | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 28 | etldb | MNO\_DAILY\_AGG\_weekday | CELL\_ID string  ,CALL\_START\_DATE string  ,NO\_OF\_EVENTS int  ,NO\_OF\_UNIQUE\_ID int | Day wise aggregstion on cell\_id level on week end | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |
| 29 | etldb | AGG\_DAY\_WISE\_7am\_7pm | CALL\_ORIGINATING\_NUMBER string ,CALL\_START\_DATE date  ,CELL\_ID string ,NO\_OF\_EVENTS int | Day wise aggregation on each subscriber at cell\_id level on day data | Day wise cell\_id level aggregation done on MNO\_7AM\_7PM |
| 30 | etldb | HOMELOCATION\_DAY\_7am\_7pm | CALL\_ORIGINATING\_NUMBER string  ,CALL\_START\_DATE date  ,CALL\_START\_TIME timestamp  ,CELL\_ID string  ,RANK int | Home location for each number on day data | For each number in a given day , we are deriving the home location, based on cell id. 1. From which tower the 'number' made most number of calls 2. When there is a tie between multiple towers , the tower from where the last call made will be assumed as Home location |
| 31 | etldb | MNO\_DAILY\_AGG\_7am\_7pm | CELL\_ID string  ,CALL\_START\_DATE string  ,NO\_OF\_EVENTS int  ,NO\_OF\_UNIQUE\_ID int | Day wise aggregstion on cell\_id level on day data | This table contains day wie aggregation on cell\_id level with no\_of\_unique records from homelocation table. |

# Reference Documents