Progress update 1/10/2015 - Fauzy Bin Che Yayah

This document is generated for the explanation on how to do the data aquisition from the original datasoure inside the Enterprise Data WareHouse (EDWH)

## Data Exploration

* Acquiring dataset for 100 records, for each zone , randomize , selective year ; ie . 2015

Below is the dataset column name :-

a <- read.csv("table\_struct.csv")  
names(a)

## [1] "tt\_row\_id" "tt\_num"   
## [3] "tt\_type" "tt\_sub\_type"   
## [5] "status" "severity"   
## [7] "important\_message" "appointment\_flag"   
## [9] "nova\_account\_name" "nova\_subscriber\_num"   
## [11] "nova\_account\_num" "package\_row\_id"   
## [13] "created\_by" "category"   
## [15] "symptom\_error\_code" "priority"   
## [17] "product" "sub\_product"   
## [19] "package\_name" "network\_tt\_id"   
## [21] "swap\_order\_num" "cause\_category"   
## [23] "cause\_code" "resolution\_code"   
## [25] "closure\_category" "resolution\_team"   
## [27] "service\_affected" "service\_order\_num"   
## [29] "btu\_type" "owner"   
## [31] "owner\_name" "group\_owner"   
## [33] "owner\_position" "btu\_platform"   
## [35] "dp\_location" "created\_date"   
## [37] "pending\_verify\_date" "closed\_by"   
## [39] "closed\_date" "source"   
## [41] "installed\_date" "description"   
## [43] "repeat\_ticket\_count" "follow\_up\_ticket\_count"  
## [45] "fdp\_device\_name" "fdp\_site\_name"   
## [47] "olt\_site\_name" "exchange"   
## [49] "timestamp" "contact\_id"   
## [51] "contact\_name" "contact\_office\_phone"   
## [53] "contact\_mobile\_phone" "contact\_home\_phone"   
## [55] "contact\_email\_addr" "due\_date"   
## [57] "part\_num" "network\_layer"   
## [59] "network\_row\_id" "asset\_id"   
## [61] "ptt" "zone"   
## [63] "service\_point\_id"

Total Zone available : 53

Air Itam,Bangi,Bangsar,Banting,Batu,Batu Pahat,Bayan Baru,Bintulu,Bukit Anggerik,Bukit Mertajam,  
Bukit Raja,Butterworth,Cyberjaya,Gombak,Ipoh,Kajang,Kepong,Keramat,Kinrara,Kl Central,Klang,Kota Kinabalu  
Selatan,Kota Kinabalu Utara,Kuching,Kulim,Langkawi,Maluri,Melaka Utara,Miri,N. Sembilan Utara,Pandan,  
Pelangi,Perlis,Petaling Jaya,Puchong,Seberang Jaya,Senai,Sg Petani,Shah Alam,Sibu,Skudai  
Pontian,Stampin,Subang Jaya,Taman Petaling,Tampoi,Tar,Tasek,Tasik Ampang,Tdi,Teluk Intan,Terengganu  
Selatan,Teruntum

## Rules for acquiring dataset

\* status = 'Closed' # dataset must be closed for complete information  
\* network\_tt\_id is NULL # not related to NTT  
\* trouble ticket type = 'PASSIVE'   
\* cause\_category , package\_name , product , sub\_product is NOT NULL  
\* installed\_date , created\_date , closed\_date is NOT NULL  
\* created\_date and closed\_date is NOT NULL  
\* length description > 10 # enough details of messages

## Sample SQL acquiring dataset from Impala

select tt\_row\_id , tt\_num , status, installed\_date , created\_date,closed\_date,tt\_sub\_type,category,  
symptom\_error\_code,product,package\_name,sub\_product,  
cause\_category,a.cause\_code,resolution\_code,closure\_category,btu\_platform, btu\_type,  
dp\_location,c.zone\_name,a.exchange , description  
from nova\_trouble\_ticket a join active\_code b on (trim(a.cause\_code) = trim(b.cause\_code)) join  
exchange\_zone c ON (trim(a.exchange)=trim(c.building\_id)) and (b.code <> 'PASSIVE' )  
where c.zone\_name like '%ZONE KEPONG%' and a.status like '%Closed%' and length(a.cause\_category) > 1  
and length(a.created\_date) > 6 and length(a.closed\_date) > 6 and length(a.installed\_date) > 6  
and a.package\_name not like '%null%' and a.product not like '%null%' and a.sub\_product not like '%null%' and length(a.description) > 10  
order by rand() limit 100

## Encoding

* Re-encoding the dataset

## Sampling

* Finding the independent variables and dependent variable.
* Sampling method
* Using the independent variables for prediction