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 Kinaba
```

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

- From Impala , loop the code , generate the SQL and replace the [**ZONE**] with the value from the zone List
- 'PASSIVE' elements - http://www.excitingip.net/53/an-overview-of-active-and-passive-components-used-to-create-an-ip-

```
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 ]%' 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%
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
-