Author: - Murari Goswami

Date: 29/05/2015

Purpose: Assignment submitted for GetYourGuide

Customer Analysis

Abstract:

Here we will analyse the customer usage pattern from the call data records generated from a billing platform. This document will give the details of the different data inputs categorized into dimension and fact tables. The idea of this analysis to get customer usage pattern and generate different BI reports from the final data cube.

Introduction:

This document will give details of chain of ETL process to load data into fact tables and populate the dimension tables. At the final product, a data cube will be generated. Various data analysis i.e. drill down, slicing, dicing can be done on this data cube and required report can be generated through it.

Target BI Report:

We expect to extract below reports from the call data record data cube.

Customer Profiling:

- ✓ Gender wise service usage classification and volumetric analysis
- ✓ Gender wise Billed Unit Volumetric Analysis
- ✓ Measure of international call data across geography

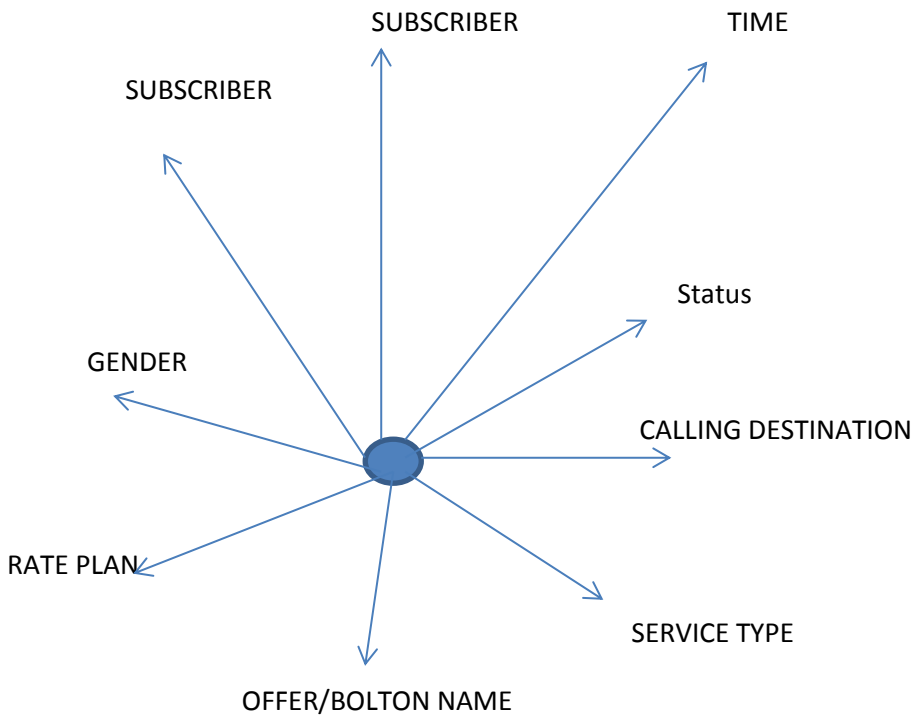
Customer Usage Pattern Analysis

- ✓ Analytic - to find Country wise- offer (Bolton Name) – Volume of Billed Units.
- ✓ Usage and revenue analytic across dimensions on service type and plan.
- ✓ Nature of outgoing calls (i.e. Fixed Landline, Emergency Number, Utility Number,)
- ✓ Rate plan analysis.
- ✓ Gender across Money spend pattern analysis

Data Modelling:

Conceptual Data Model:

At this level, relationship between data set of information can be gauged. Here given below the dot diagram for the factual information that can be retrieved from raw data while crossing this through variety of dimensional information.



Logical Data Model:

The logical data model can be viewed through a set of dimensional hierarchies in which the various dimensional data are stored. The Dimension can be categorised as :

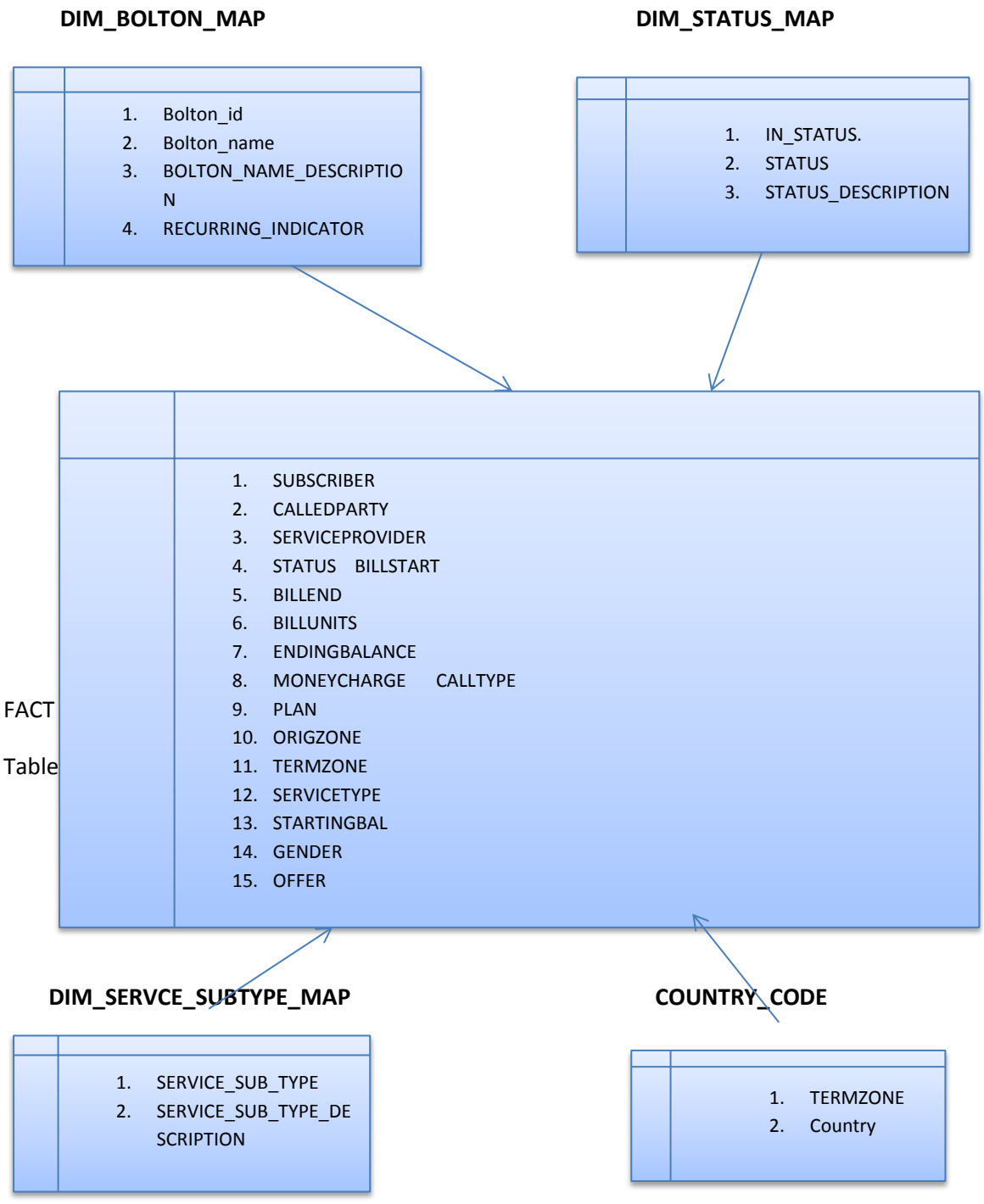
DIM_BOLTON_MAP: This dimension table has offer id that maps to Bolton name from raw CDR data.

DIM_STATUS_MAP: This dimension table has status look up details. The raw data transformation maps the status value from the reference information.

DIM_SERVCE_SUBTYPE_MAP: This dimension table holds the service tape information in it. The raw data can validate the transformation information and maps to the referential values.

COUNTRY_CODE: This dimension table look up the country code column and generate the country information for all the called party values from the raw data.

This can be shown through ER diagram as below.



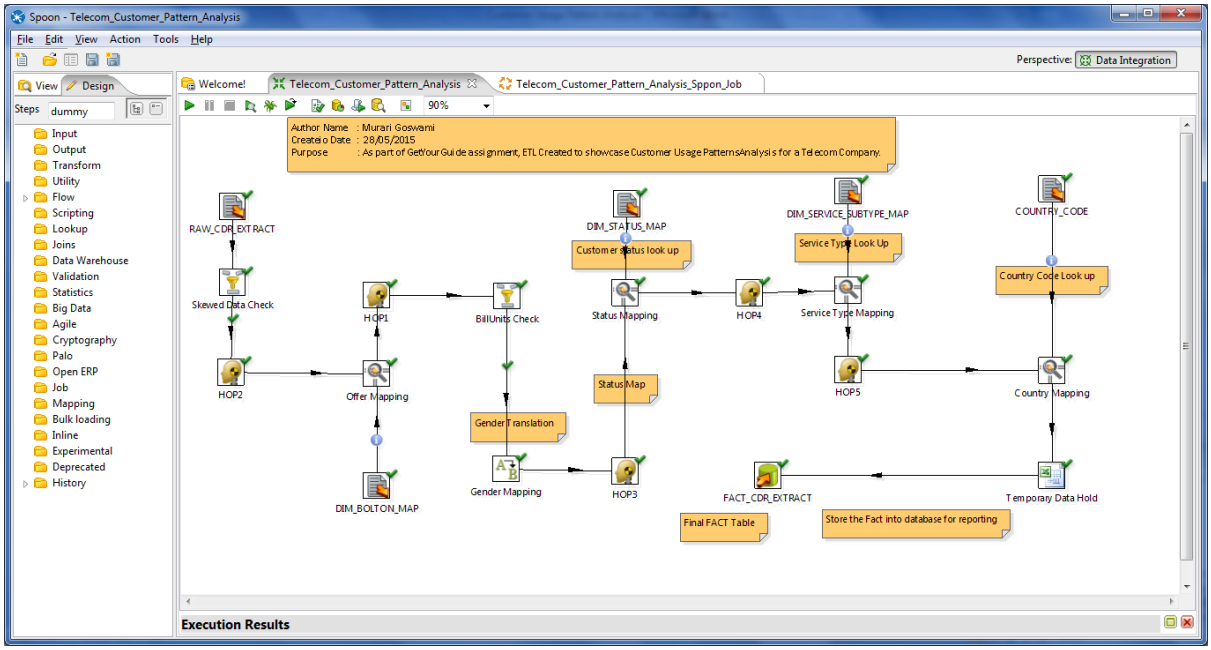
Physical Data Model and Process Flow:

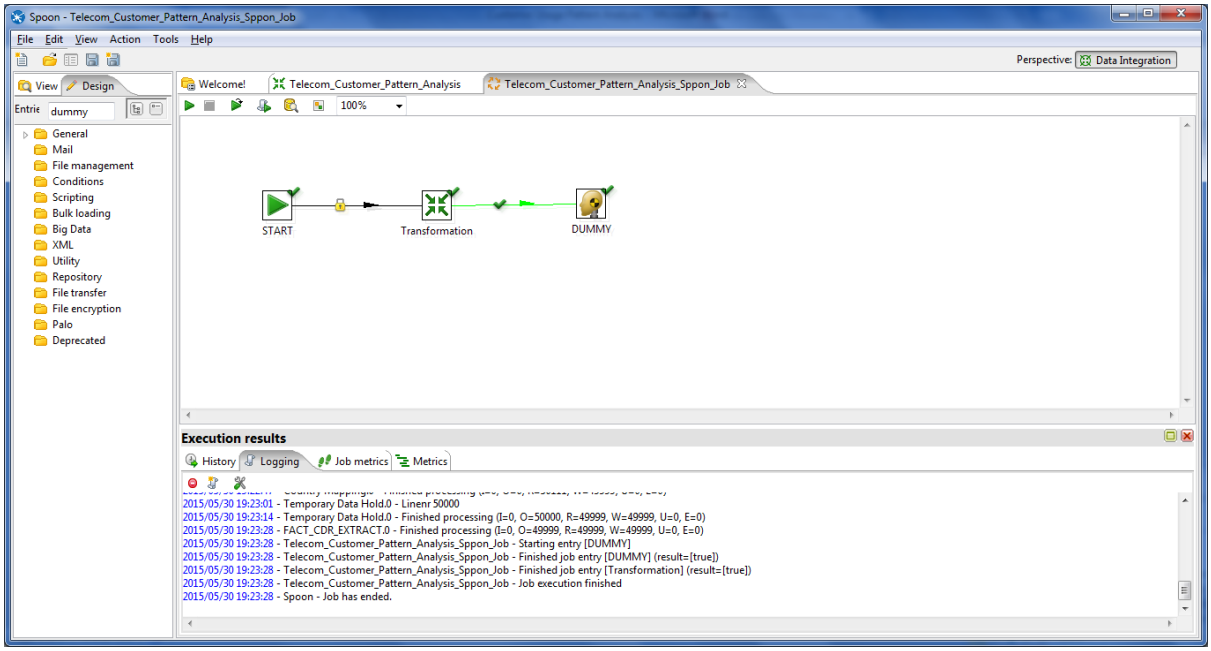
The ETL process of RAW CDR extraction, cleansing, mapping and loading into conventional FACT table to extract meaningful and useful information across different measure can give a sound foundation of information on desired need.

The steps can be broken down as per below set of data flow.

1. `The RAW CDRs will first get validated for skewed data check and erroneous data (for vital subscriber) will be filtered out at the initial stage.
2. On successful validation, the data will be hopped and then run through the mapping exercise across DIM_BOLTON_MAP to get the mapping information and additional look up column of BOLTON_NAME will be retrieved at this level.
3. At the next step of action, a translation will be done by checking the gender flag column and conversion will be done for Male and Female.
4. A hop will be there and then the next mapping will be done through look up check with DIM_STATUS_MAP. This will give the status value.
5. Taking this status value the process will now flow through Service Type Mapping by running the transformation against DIM_SERVICE_SUBTYPE_MAP table. This will give valid and required information to extract the description field that is readable at the presentation layer.
6. The Country Mapping will be done at this level where each country code will be changed to Corresponding country values.
7. At this end the data will be temporarily stored at the jump off layer and then this FACT table (i.e. FACT_CDR_EXTRACT) it will be inserted into an RDBMS.
8. I have inserted data into an Oracle database; however any database can be used.

Given Below the Spoon Transformation and job screen shots.





The design of the fact table

Once the Fact is inserted into the database, the reporting layer is made ready to extract the report as fabricated in the initial section.

All the Reports layout and report prpt files are attached in the email. Here I am sharing the SQL behind the report to get the idea of extracting the information.

Customer Profiling:

- ✓ Gender wise service usage classification and volumetric analysis

```
Select GENDER, service_sub_type_description, COUNT(*)
From FACT_CDR_EXTRACT
Where service_sub_type_description is not null
Group by GENDER, service_sub_type_description
```

- ✓ Gender wise Billed Unit Volumetric Analysis

```
Select GENDER AS Gender,
    Sum(billunits) AS CALL_VOLUME
From FACT_CDR_EXTRACT
Group by GENDER
```

- ✓ Measure of international call data across geography

```
Select Country As Country , sum(billunits) As Volume_Usage
From FACT_CDR_EXTRACT
Group by Country
```

Customer Usage Pattern Analysis

- ✓ Analytic - to find Country wise- offer (Bolton Name) – Volume of Billed Units.

```
Select distinct country as country,
                        bolton_name offer_name,
                        sum(billunits) over (partition by bolton_name) as vol_bill
From fact_cdr_extract
Where country is not null
```

- ✓ Usage and revenue analytic across dimensions on service type and plan.

```
Select distinct service_sub_type_description as service,
        plan,
        sum(billunits) over (partition by plan) as vol_bill_plan,
        sum(moneycharge) over (partition by plan) as rev_plan
From fact_cdr_extract
Where country is not null
Order by rev_plan desc
```

- ✓ Nature of outgoing calls (i.e. Fixed Landline, Emergency Number, Utility Number,)

```
SELECT (SELECT SUM(BILLUNITS)
          FROM FACT_CDR_EXTRACT
         WHERE CALLEDPARTY LIKE '1%') AS LAND_LINE,
       (SELECT SUM(BILLUNITS)
          FROM FACT_CDR_EXTRACT
         WHERE CALLEDPARTY LIKE '2%'
            OR CALLEDPARTY LIKE '3%'
            OR CALLEDPARTY LIKE '4%') AS PREMIUM_NUMBERS,
       (SELECT SUM(BILLUNITS)
          FROM FACT_CDR_EXTRACT
         WHERE CALLEDPARTY IN ('901', '905', '906')) AS EMERGENCY_NUMBER,
       (SELECT SUM(BILLUNITS)
          FROM FACT_CDR_EXTRACT
         WHERE CALLEDPARTY LIKE '8%'
            OR CALLEDPARTY LIKE '9%') AS UTILITY_NUMBER,
              (SELECT SUM(BILLUNITS)
          FROM FACT_CDR_EXTRACT
         WHERE CALLEDPARTY LIKE '7%') AS MOBILE
  FROM DUAL
```

- ✓ Rate plan analysis.

```
Select plan, sum(billunits) vol_unit, sum(moneycharge) mon_charge, sum(endingbalance-
startingbal) diff_charge
From fact_cdr_extract
Group by plan
Order by vol_unit desc
```

- ✓ Gender across Money spend pattern analysis

```
Select gender, plan,sum(moneycharge) as charge
From fact_cdr_extract
Group by gender,plan
```

PLEASE SEE BELOW.

INPUT DATA:

This is the assumption that the input files are stored @ -

C:\Personal\Yougotoguide\HomeWork\Input Data\CSV.

The database connection was used to store the FACT table and then all reports SQL are filed on that FACT table.

There is no other dependency with the database. Hence in order to test the files, it is recommended to store the FACT table in the database an while doing to please use any of the database connections. I have used jdbc thin client connection.