Implementing MDM (with PME) for Airlines Industry

Joydev Mandal

Table of Contents

1. Introduction 2
   - 1.1 Purpose 2
   - 1.2 Scope 2
   - 1.3 Intended Audience 2
   - 1.4 Overview 2

2. Goals 3

3. Benefits 3

4. Key Challenges 3

5. Solution Outline 4
   - 5.1 System Context 4
   - 5.2 Solution Architecture 4

6. High level design details 5
   - 6.1 Customer Definition 5
   - 6.2 Logical Entity Model 6
   - 6.3 Component Model of MDM integration parts 6
   - 6.4 Some of the key Probabilistic Matching Engine (PME) design strategies 7
     - **6.4.1 Bucketing Strategy for PME** 7
     - **6.4.2 Defining matching threshold** 7
     - **6.4.3 Defining anonymous values list** 8
   - 6.5 Some key architectural and design decisions and recommendations 9

1. Introduction

1.1 Purpose

Purpose of this document is to provide a set of pointers and guidelines on why and how to implement Master Data Management for Airlines industry along with Probabilistic Matching Engine enabled.

Master Data Management for customer data has become a basic requirement in any dynamic industry like Airlines for better customer insights. And this insight in master data is also helping to the next level of analytics and campaign management using the transactional data with identified customers. This combined usage of master data and transactional data is going to give more business value to the client with more optimized campaign program and other analytics reports for defining better marketing strategy.

In this context, the document provides the information on such kind of engagement and implementation pattern in airlines industry. It also provide some valuable information about the solution architecture, logical model, some key PME design decisions and some overall architecture and design decisions which will help any other team in similar engagement.

1.2 Scope

Scope of the asset is architecture and design of such project having requirement of Master Data management with Probabilistic Matching Engine in Airline Industry.

1.3 Intended Audience

It is intended for Solution Architects, designers who are involved in any master data management engagement using Probabilistic Matching Engine in Airlines Industry.

1.4 Overview

This document can be used to understand business goals, challenges, expected benefits out of this kind of engagement, solution architecture, system context, component model and PME design strategies for customer match and merge mechanism and finally some key architecture and design decision and recommendation for such kind of implementation.

2. Goals

The goal is to create a customer Master Data Management application within its landscape, so as to have a consistent view of customer profile (individual and organization) across various functional and business units.

3. Benefits

Main objective is to set up a customer master data management application to track their frequent flyer base (say around 30%) and their non-loyalty members (say around 70%) in a centralized customer data management application. The advantages of having a centralized customer data management are:

- The value obtained from a loyal customer is higher than from a non-loyal customer. Typically a frequent flyer member books using the airline portal rather than from GDS (Global Distribution System). This effectively means there is a huge cost saving from moving from a GDS assisted booking (cost around 8 USD) to an airline hosted booking (cost around 40 cents) for an airline perspective.
- Ability to take trusted decision based on the consistent view of customer profile information, across business units.
- Understand the non-loyalty members as well, who has provided the sufficiency on the data elements as stated in the client requirements, to help move those customers into loyal customer base.
- Utilize the profile class details to help passenger in the value chain move up the gradation of non-loyal profile class to alliance profile class and finally to Member profile class.
- To manage the preference and consent information in a centralized manner using this customer master data.
- To send more optimized and target oriented campaign program to identified customers to get more business in a more cost effective way.

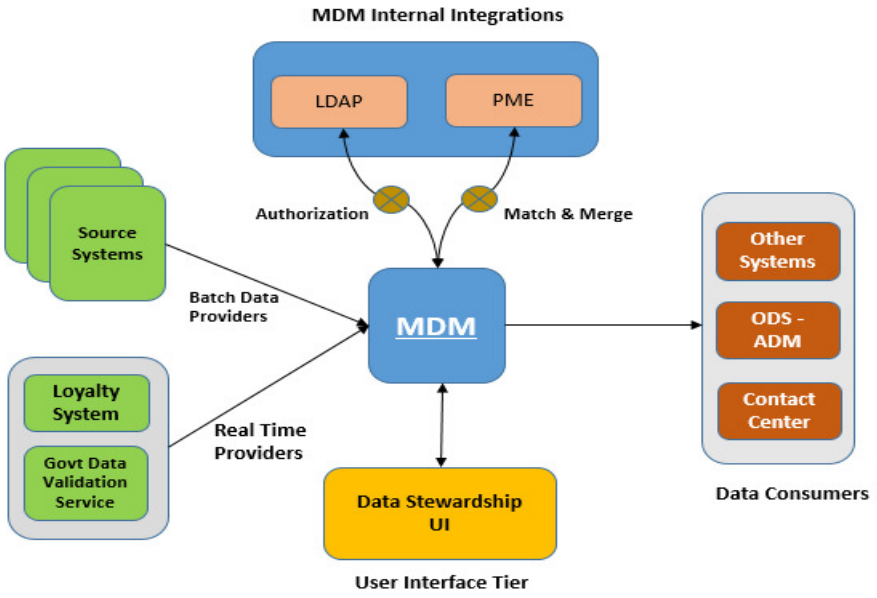
4. Key Challenges

The key business challenges for such an implementation in Airlines Industry are:

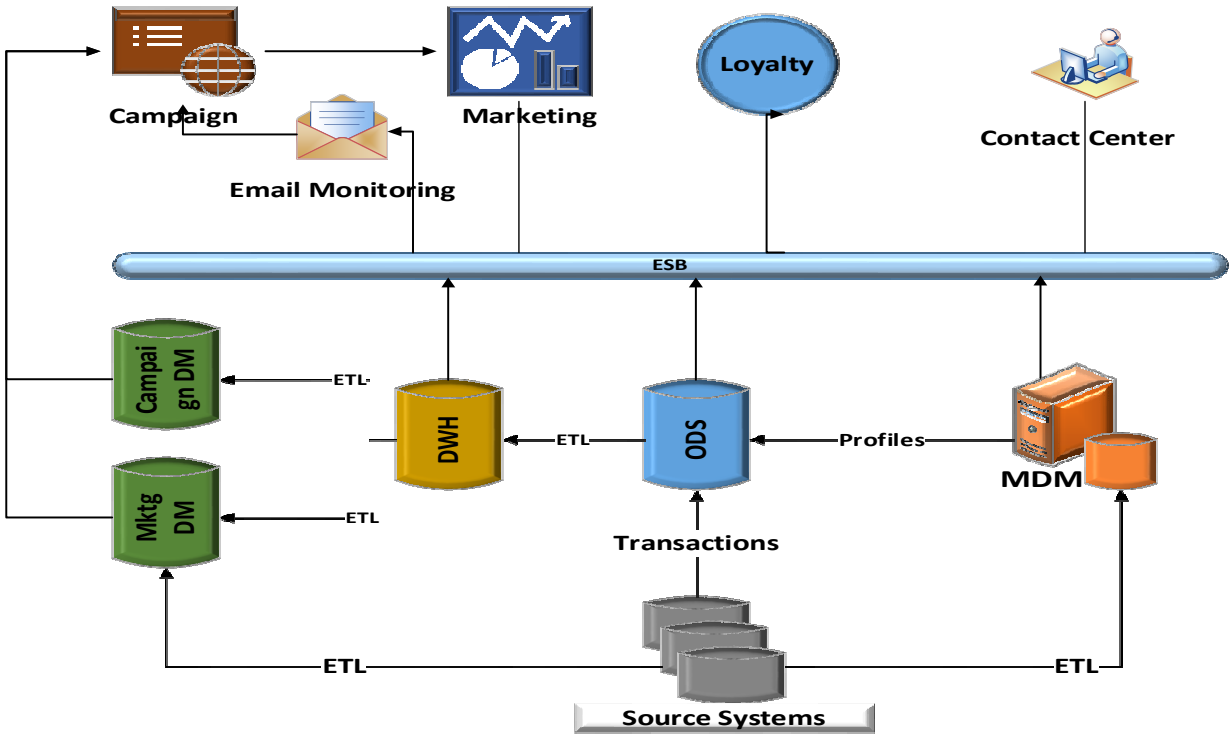
- Isolated organization, with very limited collaboration across lines of business.
- Data is still predominantly residing within lines of business.
- No regular treatment of data (quality) related issues on the lines of business data
- No enforced KYC system in this industry
- Not having any enterprise level business key for its customers identification
- Identifying potential customers from day to day transactional information and merging them into master data bases because of very limited unique information specially for domestic travel
- No scope or reference available for key data standardization like Address, Name etc. This causes the probabilistic matching to be really challenging.
- Decision makers & business analysts depend on ad-hoc reports that are not consolidated. There is no consolidated data governance system across lines of business.

5. Solution Outline

5.1 System Context



5.2 Solution Architecture



6. High level design details

6.1 Customer Definition

Recommendation for a profile information of the individual or corporate customers, should have;

- ✓ Identification information (TR Identification no, Passport no etc.)
- ✓ demographic information (Name, Address, DoB, Gender)
- ✓ Communication information (telephone, e-mail etc.)
- ✓ IDs in integrated systems and recorded with unique ID.

Recommendation is to use all available attributes (keys) valuable in matching duplicate party records in PME. The resulting 'match category' counts, and more importantly, level of accuracy in matching and searching will be more when more attributes are present for PME to use.

In general the key attributes or the critical information are as mentioned below:

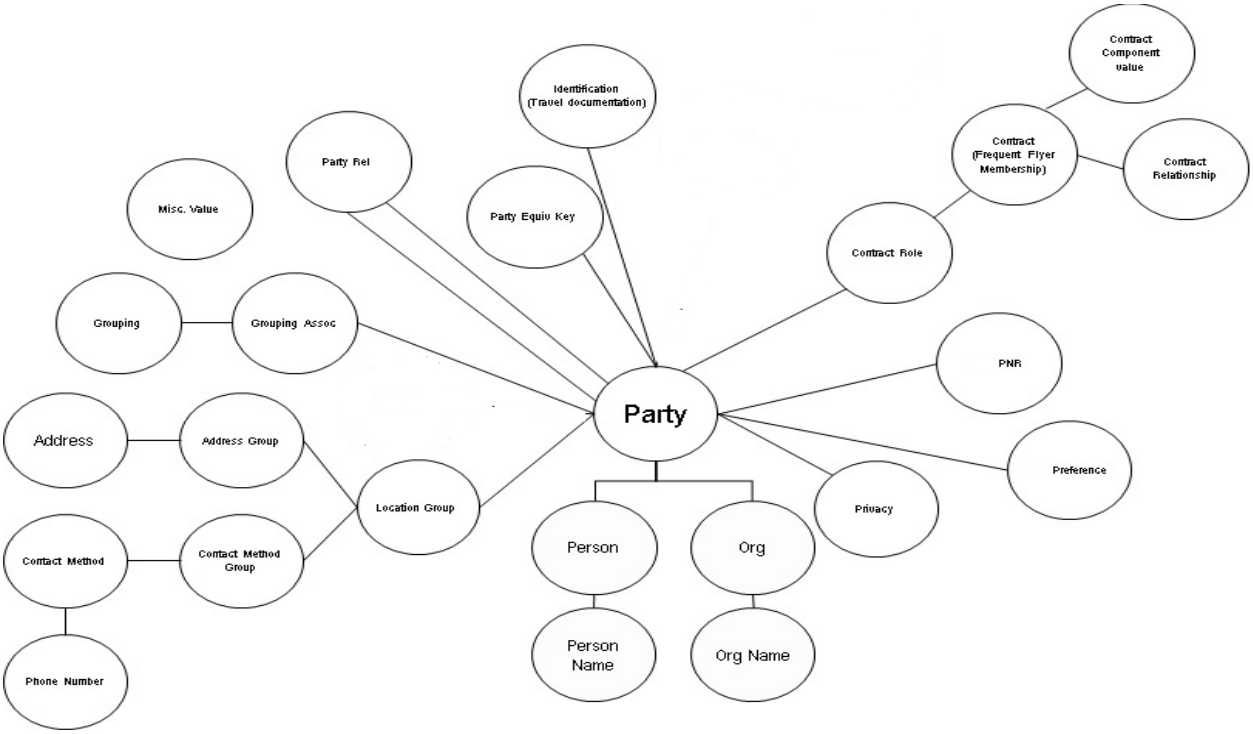
#	Attribute Group	Attribute	Remark
1	Identification	Government Id (like SSN etc) Passport Number Frequent Flyer# Any other organization specific business key	This category of field is critical as it helps to manage members (Frequent Flyer holder), non-members (alliance member and identified individuals) and identified prospects (identified individuals)
2	Communication	Telephone number or Email address	None
3	Demographics	Address – Preferred Gender Title First name, Last Name Date of birth	None

Now the challenging part is identifying a profile from PNR or reservation point of view. To create a profile, customer definition (profile information) could be achieved under following criteria.

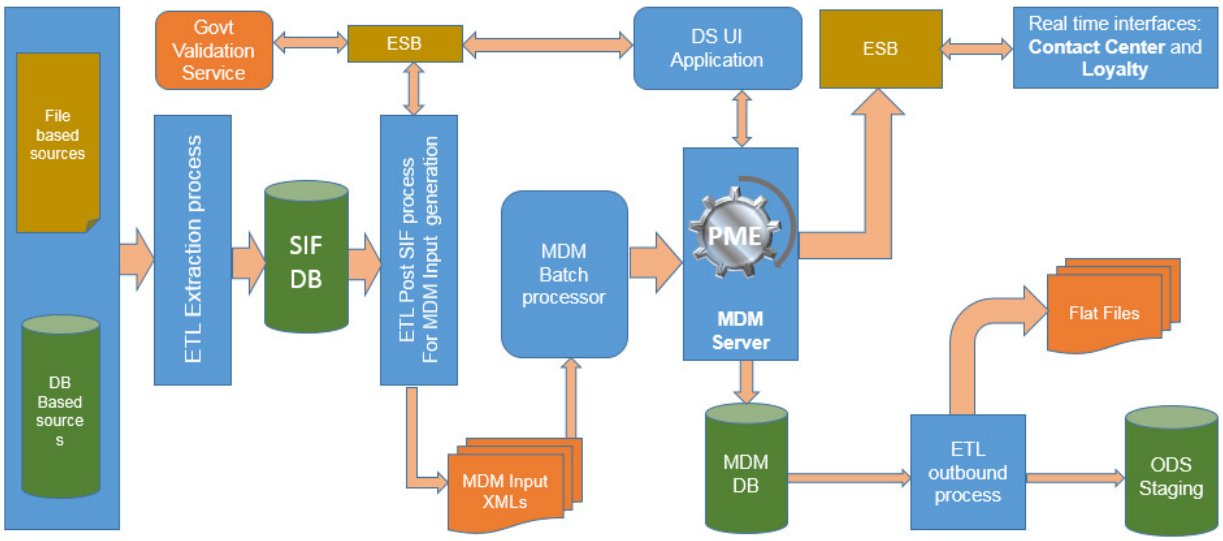
- Identification Information:
  - Passport numbers: Mostly present for the international traveler. Very limited availability for domestic passengers
  - Frequent Flyer number where ever available during the reservation
- Communication information: (telephone, e-mail, address) may not be always passenger related. And maybe just agency communication information. So, extensive list of these agency email information are to be included in the anonymous value list for PME.
- Demographic: This may include demographic information in rarely formatted/ unformatted way. So, name and address standardization needs to be enforced before accepting the data inside MDM for matching.

As described above, information needed for customer definition will be rarely found together in a formatted way. As a result, creating profiles based on this criteria may be problematic or rare if not merged or synchronized with other sources. So except unique system ID, all other information is either rare, optional or unformatted. To resolve the issue, recommendation is to use business keys to match the reservation data with other transactional sources like departure control system etc by using business keys to make the sufficient profile information. The rules are to be defined in the as is process workshop with client for reservation / transaction based source systems.

6.2 Logical Entity Model



6.3 Component Model of MDM integration parts



6.4 Some of the key Probabilistic Matching Engine (PME) design strategies

6.4.1 Bucketing Strategy for PME

Buckets are a concept used by PME to logically group records together to conduct more significant comparisons (*compare only those records w/ similarities in data*) and to increase performance.
The goal of bucketing is to speed up the matching process by grouping specific attributes and further group 'like' values within those attributes so that only those records sharing similar attributes are later compared to one another
Bucketed attributes help the engine find other records that share similar data in common.

These buckets are based on best practices which are compiled over many projects and overall experience in the field. Other than lowering the number of records to a smaller, more significant (*similar data*) set of records for increased efficiency and performance, PME buckets are also meant to be configured so that the target record(s) is found for comparison or searching. Meaning that the probabilistic functionality available to PME (*nicknames, phonetics, edit distance, etc.*) will also help widen the list of records to increase the chances of finding that target record(s). This is the balance for which these recommended buckets based on 'best practices' will achieve. To remove, change, or add to these listed buckets may create:

- Redundant buckets which will yield the same records with already existing ones
- Unnecessary clutter in underlying DB tables and decrease performance
- Groupings of records which are too small to find the target record to compare against or searching for

Typical Buckets are:

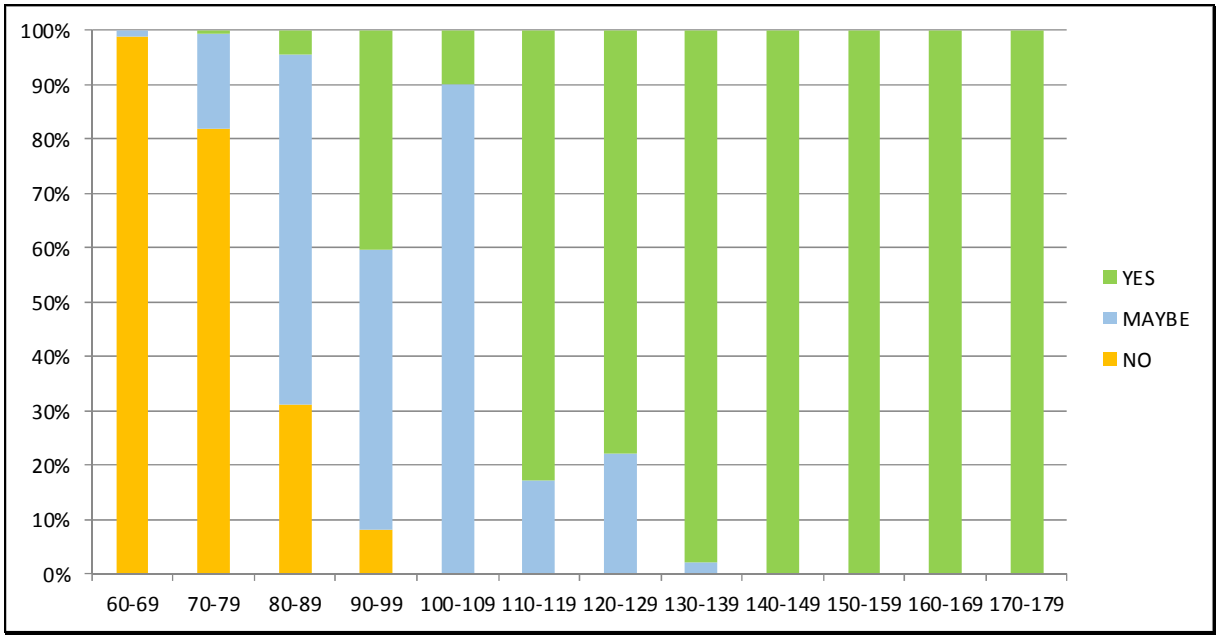
- Name Only (*at least 2 names: first + last*)
- Name + Phone (*still to be configured and to replace NAME + Postal Code*)
- Name + DOB (*at least 1 name + date*)
- Phone Only
- TCKN ID Only
- Passport Only
- FFP Number Only
- Email Only
-

6.4.2 Defining matching threshold

There are two types of threshold to be decided during the PME workshop after analyzing the weight calculation and sample pair generation process.

- **Upper/Auto-link** threshold reflects the score at which you are confident that a match represents the same person. We set this based on the tolerance for false positives.
- **Lower/Clerical Review** threshold reflects the score at which you want to manually review matches. We set this based both on the tolerance for false negatives and on the number of matches client willing to review, based on the cost of having people manually review these tasks.

Sample outcome of this type of analysis to define the threshold as mentioned below:



Considering the above analysis report as one sample, recommendation should be as mentioned below:
One conservative and one aggressive threshold sets. Delivering these two sets of thresholds not only maximizes PME's value specifically to client, but also gives them an idea of what the results may be if the final decided thresholds are slightly above, below, or in between the suggested recommendation. Also regardless of any positives in 'match category' counts, these thresholds reflect the answers client recorded in the 'sample pairs' they normally review. (***Highest NO & MAYBE*** = *~10.0* / ***Lowest YES*** = *~7.0*)

Conservative	**Aggressive**
AL(Upper)=11.5 CR(Lower)= 7.0	AL(Upper)=10.0 CR(Lower)= 8.0

6.4.3 Defining anonymous values list

The items discussed below talk about PME's treatment of invalid, fake, or anonymous data, PME does come with a set of anonymous values 'out of the box' however, said list can be further enhanced by adding anonymous values which are custom to client data. Said anonymous data can be flagged by PME as such and appropriately not include in the matching which PME conducts. Such a value (*for example: an anonymous PHONE*) will be treated as if it's a NULL.
Recommendation to client is to update the list of anonymous values with any not already captured by 'out of the box' values, or initially found w/ high frequencies or 'sample pair' reviews and send back to include in custom PME.
Recommendation is that the list of rules laid out by Client to consider in determining an anonymous value to not be accounted for in PME rather, in some other area (*perhaps before coming into MDM: like ETL*) of the greater MDM architecture. There are way too many possibilities for anonymous values to keep as static values in the packaged and deployed PME. Best to keep in an area like ETL where client business can dynamically add to such a list and prevent from ever being loaded into PME.

6.5 Some key architectural and design decisions and recommendations

#	Domain	Type	Decision
1	Customer data sourcing requirement	Functional	Passenger data which has sufficient amount of data will be managed inside the Centralized Customer database. Sufficiency of data relates to as indicated below. • Mandated element such as first name and last name, title, date of birth, gender etc. – individual traveler • Mandated elements such as organization first name, demographic address and Taxation number etc. • Mandated element such as PNR data and PNR creation date • Mandated element such as Email address, social media, and Telephone details • Address details • Travel documentations and/or Government identifier key provided for local travelers. • Any passengers with group flyer identifiers from any Alliance, own Frequent Flyer program will also be considered. • Cross reference keys from various systems • Privacy and legal terms agreeing that client could contact passenger through campaigns etc, for service fulfillment and/or digital convenience. • Preferences, car, auxiliary, hotel, destination, leisure etc • PNR with multiple passenger in a single PNR and Group PNRs where all the four identifying elements required for profile are provided only. Passenger data with active PNR for a specific period (say 3 years) with valid booking or all FFP cards will be loaded
2	Architectural pattern	Technical	Recommendation is that the MDM application be introduced within the client's existing landscape as a consolidation pattern. This way all the changes to the master data records e.g. PNR data, loyalty profile data, corporate data, campaign data happens in front office application. The changes are then consolidated into the MDM and ODS for view and for analytical usage. This way it may not be required to change core application to subscribe to master data updates or creates in real time, making them MDM aware and web services enabled. This pattern helps to plug in MDM into the landscape in a faster manner and improve operations, in addition to help business take trust based decision.
3	Common canonical or Standard interface	Technical	It is recommended to use a SIF landing area ahead of the ETL tier, so as to do the following. Parse the data and store this as 2 NF form, enrich the data and standardize in the relational table. Identify within the SIF

	format landing area		possible updates or add based on cross reference keys to the Enterprise customer identifier keys. This way the output XMLs are of similar structures rather than composite XMLs Retain the data for a period of 3 months to help reconcile source data issues Utilize the SIF landing area to identify data anomalies and report this back to source system, using proactive reporting mechanism. Ensure there is a standard data contract between client's source system and MDM, where by new data domains and new field types are detected upfront. The data model of SIF is target system based i.e. based on the MDM data load, so that the data type and data length is proven and could be used for web services integration in the future.
4	Utilizing XML for ongoing loads	Technical	All data additions and data extension will re-use the native XML for loading. Additional the same native XMLs will be used for ongoing loads. It uses the ETL to extract the data and provide in pre-defined XML formats. One common canonical format will be used for integration and data migration from upstream application to the MDM server. In addition to this, the variation may be on Organization and Person entity.
5	Probabilistic match engine instead of deterministic match engine	Technical	It is recommended to go ahead with Probabilistic match engine, so that the scoring is fine grained and it helps the data stewards to consolidate the data better. With the newer version of the Infosphere MDM server v11.3 it is possible to create buckets and then compare tokens, there by assign weights which are summed together to arrive at a match score.
6	Rules of mastering	Technical	Rules of mastering to be defined to minimize the attributes that are being managed in the hub. This is very critical to minimize the attributes being stored in the MDM application and to make sure that very frequently changing attributes doesn't get qualified as master data.
7	Call center integration	Technical	The customer profile information is replicated in near real time between MDM and ODS. Contact Center uses the web services from MDM.
8	Cleansing and standardization	Technical	This step is very critical considering the Probabilistic Matching Engine is being used in the engagement. Recommendation is that cleansing and standardization be done for 30% of the core fields. The cleansing to be done must be extremely critical for the subsequent matching purposed in PME. This be limited to the following, Name, Surname, Address, Phone Number, Email will be standardized as part of the cleansing functionality.
9	Future Extensibility	Technical	Integration of MDM through ESB is recommended for systems such as Contact Centers and Loyalty. This ESB could be used for any future system integration also.