Enabling Cross Integration of Distributed Applications By Applying MDM Framework

N. Mary Vijaya Nirmala M.Tech
Associate Professor
Department of Computer Science
VNR College of Engineering
Ponnur, Guntur, Andhra Pradesh
Nirmala.neelam@gmail.com

Khaja Vali Shaik M.Tech
Department of Computer Science
VNR College of Engineering
Ponnur, Guntur, Andhra Pradesh
khajashaik.dw@gmail.com

***Abstract* - Organizations are using various platforms to store the data, depending on the need and cost, like third party hosted applications, on the cloud applications, legacy systems, data warehouse applications, Big data applications. To make it more complex web of applications hosted on different platforms, the term Big Data has been evolved in the recent years. The organizations have been facing the challenge of getting the single view of business data from hundreds of data entry terminals, in the distributed environment. This can be resolved by using Master Data Management methodology. It fixes the data quality problem on the operational side of the business. It augments and operates the data warehouse on the analytical side of the business. One analyst reports "75% of leading companies are incapable of creating a unified view of their customer" [1]. This is because; the Master data consists of facts that define a business entity, facts that may be used to model one or more definitions or views of an entity. Entity definitions based on master data provide business consistency and data integrity, when multiple IT systems across an organization (or beyond) identify the same entity differently.**

Integrating MDM framework into the Distributed Ecosystem and refining the domains of MDM using master data elements will yield excellent results to the organizations. It will help the organizations to enable cross platform integration, which will also work as closed loop feedback system and induce the data standards in all the systems of an enterprise. It will help to get maximum Return on Investment (RoI), from applications spanned across the organization and to deliver high quality of services. This paper discusses and focuses on enabling cross platform integration in distributed Eco system of an enterprise by applying MDM framework and to derive more trusted information on the foundations of truthful information.

***Index Terms*—Big Data, Master Data Management (MDM), Internet of Things (IoT), Return on Investment (RoI)**

I. Introduction

The increasing amount of data is creating challenges to organizations in Eco system of distributed architectures, which are very common in companies now-a-days. Workgroups, such as organization departments, develop data processes in silos which lead to variance in the business concepts and object definitions. The need to share information across the organizations and supply chains is driving data from silos to be exposed, unified and shared. This reveals enormous data discrepancies and incompatibilities.

Master Data is the critical business information related to the transactional and analytical operations of the enterprise. Master Data Management (MDM) is a combination of applications and technologies that consolidates, cleans, augments corporate data and synchronizes it with all applications, business processes, and analytical tools. This results in significant improvements in operational efficiency, accurate reporting and strategic decision making. Master data describes the business-oriented properties of data objects which are used in the different applications across the organization together with their associated metadata, attributes, definitions, roles, connections and taxonomies. Master data is the data that has been cleansed, standardized and integrated into an enterprise-wide system and used across multiple business domains.

One of the best examples to understand the importance of MDM framework with big data is, analyzing the streaming information generated out of rental car sensors to find out simple truth – is this customer a safe driver? Depending on the driving pattern of an individual for the duration of rent (sensor data analysis in this case), the rental company can rate the customer on a scale of 1 to 10 and store it in MDM. [2]. This information would be made available to the other applications by the MDM system.

II. WHAT IS MASTER DATA MANAGEMENT?

Master data is the high-value, core information used to support critical business processes across the enterprise – information about customers, suppliers, partners, products, materials, employees, accounts and more – and is at the heart of every business transaction, application and decision.

Master data, therefore, is a subset of big data and master data management can provide a compelling starting point for big data analysis. MDM, by definition, focuses on the highest value entities within an organization. The goal of MDM is to streamline data sharing across systems and provide everyone in the system with a single, consistent view of critical data by using both technology and data governance techniques.

III .EXISTING SYSTEM ARCHITECTURE

Many organizations are suffering from sub-optimal customer relationships due to misaligned customer data that is kept inconsistently and siloed across different applications. Redundancies and quality issues are commonplace.
This results in:

- Reduced sales effectiveness
- High marketing costs
- Suboptimal customer service
- Unreliable analysis

All the systems have their own implementation and perform designated functions of their own. These systems are up to date in their own context and consolidated information is confusing. These systems are maintaining data as per individual system owner's standards. It is challenging, as and when there is need to integrate or sharing the data across systems. A single view of common data elements, used to integrate systems is not available due to inconsistent data standards and resulting in data redundancy. Integrating different systems available as of the day will result in the architecture shown in Fig1.

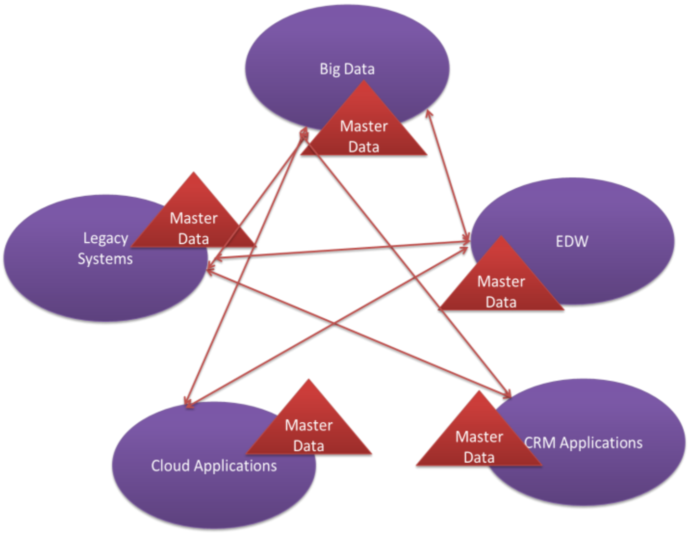

Fig 1: Architecture of different systems with their own individual MDM

No comprehensive system contains the single view of Master data. All the systems are correct and up-to date in their own context. But when consolidating and looking at the information it results in inaccurate information. As the master data elements are being maintained by all the systems, it is resulting in redundant data. But no single system is able to provide all the updated and most recent. This is the point where the need of MDM came up to maintain master data elements in a centralized repository.

The increasing complexity of data management environments and increasing demand for cross-function data exchange has created strong need for the MDM programs. The success of an organization is closely dependent on quality, consistent data and the mechanism to share the right information at the right time.

IV. PROPOSED SYSTEM ARCHITECTURE

Master Data Management is an enterprise strategy that treats master data as an asset with enormous top-line and bottom-line impact. It facilitates data consistency across multiple systems for streamlined business processes and enterprise reporting while ensuring end-to-end data stewardship and master data governance [11]. MDM is referred as a single source of truth for everything related to core information, which would be referred as data domains. Example would be customer domain, location as a domain. But will we ever be able to know 'everything' about customers given lot of data related to them today is huge and every system have different updated version of data.

In spite of such support for data integrity, enterprises had duplicates in their master data that resulted in inaccurate results in analytics on that data. For example, an enterprise may target an expensive advertisement campaign for a new product to its existing customers; however, due to the fact that a particular customer may exist with different IDs across multiple systems, the enterprise may be sending its campaign materials to the same person multiple times.

Issue in the existing system architecture is the master data elements are not being maintained centrally. Data sharing and communication between existing different systems is not centralized. Each system has its own version of information. Each system says the data being maintained by their system is true. When looking from the top, the consolidated view is not available and hence resulting in in-consistency, redundancy, duplicity and un-trustable data. There need exists, where this data require to be captured and maintained in a centralized repository, which needs to be administrated and updated properly. Architecture of the System needs to be changed as shown in Fig 2.

The approach to building the proposing this system requires three major steps:

• The first step is to assess the current mastering capabilities. During this step, one should assess the MDM maturity of the entities in scope. In order to quantify the impact of the MDM technology, it is important to have a relative point of comparison.

• The second step involves envisioning the future data mastering capabilities and the solution footprint to support them. In addition, it is important to define the implementation plan and understand the cost of implementation. This is required to define the investment required.

• The third step understands the benefits of the MDM technology to the business. It is during this stage that we quantification is done.

Using the investments from step 2 and the quantified business value in step 3, calculate the ROI for MDM.

Centralized hub would be maintains the data by following the steps as

Data Consolidation: Extract Transform Load (ETL) from source systems including data cleansing based on Data Services.

Data Cleansing: Detect and cleanse inaccurate records

Data Correction: Continuous quality control on the central MDM hub.

Master Data Centralized Hub: At data creation in, new records are matched against the existing records in the system (local check) and against the records in the MDM hub (global check). If matches are identified, the user is notified of matching records, preventing duplicates from being created.

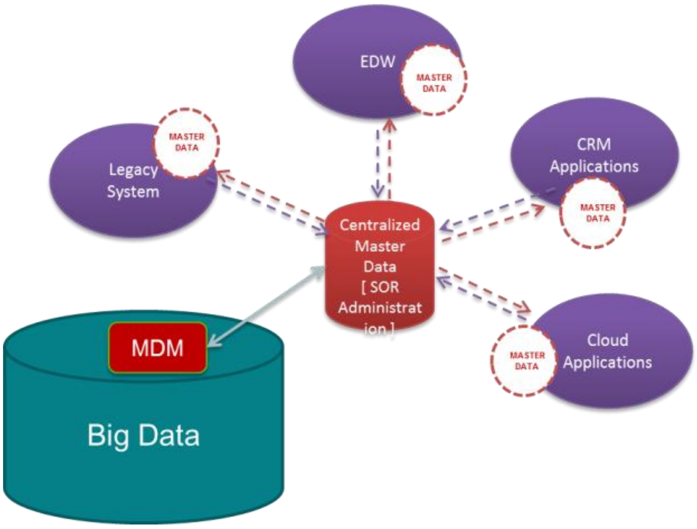

Fig 2: Architecture of different systems integrated based on Hub based MDM system.

The central repository contains what data is available and where it is stored.

Following process would help to achieve Centralized Master data Management system

- Identify System Of Record (SOR) across different Systems
- Agree Upon Data standards
- Define Data Rules to create/Update SOR
- Automate the process of data synchronization between System of Record and data subscribers/users of system
- Designate Data Stewards and Data Trustee to administer Master data

V. METHODOLOGY

V-A) Data Profiling

The first step in any MDM implementation is to profile the data. It provides the ability to understand data, highlighting key areas of data discrepancy; to analyze the business impact of these problems and learn from historical analysis; and to define business rules directly from the data [9].

V-B) Data Cleansing

Data cleansing involves standardization, error correction, matching, de-duplication, and augmentation of the data. This process detects the incomplete, incorrect, inaccurate, irrelevant parts of data based on data rule and then replacing, modifying or removing the bad data. The values rejected by a data rule are moved to error table where the cleansing strategies are applied.

V-C) Data Rules Definition

- Remove the white spaces, control characters (enter and back space) from first name and last name
- If first name concatenate with the last name is the First name for another record, and then both are same records.
- If previous record does consists of normal value and the latest record is empty then copy all previous record values to latest record
- (Eg) Old Record: First Name with enter character – Phone number is available for this record.
- Latest Record---Phone no is not available
- In the above case, copy the phone no from old record to latest record
- If any of two records have the same first name, last name, dob then they are same records
- If any of two records have the same First Name, Last Name And Email Id then they are Same Records

- If Date of Birth (DOB) differs for two People with the same email id always consider first dob entered in application systems.
- If legacy and CRM have Two DOB for same person then consider legacy CRM to update Master record.
- If all source systems contain different phone numbers for the same person, then latest phone number will be considered to update from Master repository.
- If legacy and Cloud have two DOB for same person then CRM has to be considered to update DOB in Master repository.

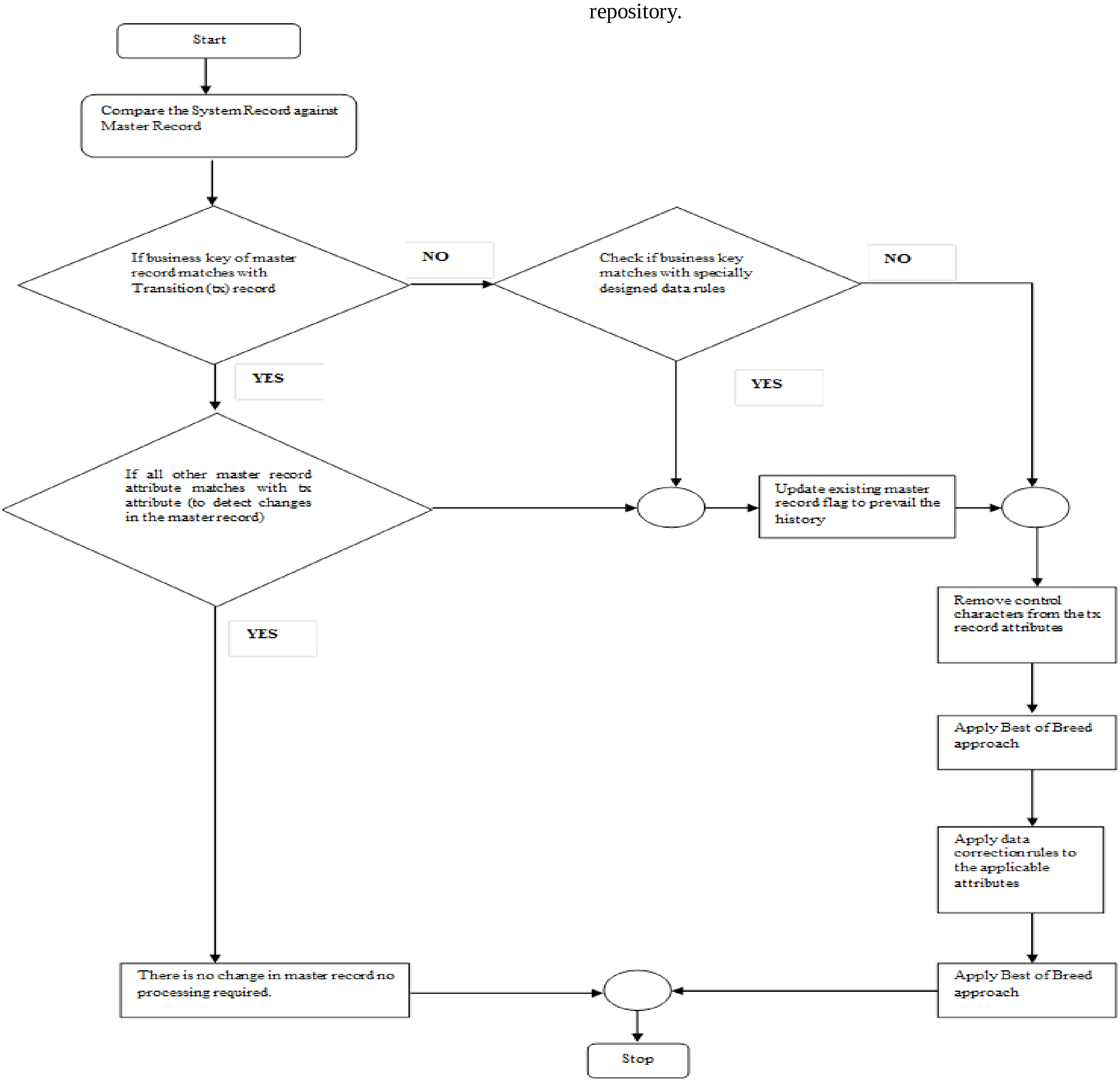
Start
Compare the System Record against Master Record
If business key of master record matches with Transition (tx) record
NO
Check if business key matches with specially designed data rules
NO
YES
YES
If all other master record attribute matches with tx attribute (to detect changes in the master record)
Update existing master record flag to prevail the history
YES
Remove control characters from the tx record attributes
Apply Best of Breed approach
Apply data correction rules to the applicable attributes
There is no change in master record no processing required.
Apply Best of Breed approach
Stop

Fig 3: Flowchart

If legacy and CRM have two different addresses for same person then address from legacy is taken as to update Master repository, if the modified date of legacy is greater than the CRM.

V-D) Data Integration

Data integration process involves populating the hub database with data from the source systems initially and keeping the source systems synchronized with the hub database as the source systems make changes to the data [10].
Check for duplicates- This process is the heart of most MDM systems. It is both the hardest and most important part of populating the MDM hub. As single view of customer data is required, records describing the same business entity must be combined into a unique record for each unique entity; Duplicate checking process will check for things like alternate spellings and missing words—for example, Raja Rajan, Mr. Raja Rajan, Rajan Raja so forth.

Load the MDM hub database- MDM hub has been loaded by following the steps shown in the Figure 3.If the new record is not already in the hub database then insert the data into the correct tables. But if it is a duplicate, the load process must check the business rules for this entity to decide, what data to update with the incoming record. This process has been detailed in Figure 3 as an example, if there is no address in the current record and the incoming record includes an address, the address is added. If there is already an address and the incoming record also has one, there must be a rule specified to decide which one to keep or if both should be kept. If the business rules can't resolve the conflict, the incoming record should be put on a queue for manual processing. The key of the record should be added to the database to record the connection from the hub record to the source record. This may be used by queries to find the source record or by the hub to publish hub updates to the source systems.

Update the source systems- If loading a new record changes the hub database, the change may need to be propagated to one or more of the source systems. For example, if a new, address is added to a customer record, other applications that stored information about that customer may want to use the new address.

The load process works best if the most authoritative complete data sources are loaded first, so that subsequent loads make relatively few changes to the existing hub data. Primarily, however, it's best to record duplicates and synchronize the application data with the hub data. Loading the most critical databases first also leads to earlier time to value.

V-E) Data Synchronization Process

Synchronization is a process that transfers changed master data records from the source application that made the change to the MDM hub [11]. This introduces the possibility of conflicting updates and inserts from multiple systems, and it introduces some latency between the time a change is made and when it shows up in the MDM hub database; so the business must understand the limitations of this system. After a change has been detected in the source system, it should be sent to the MDM hub as quickly as possible to reduce the update latency. So the frequency of updating master between hub and source applications should be too close to maintain and make available master records in real time.

V-F) Best-of-Breed Approach

Best-of- Breed is a set of different modules collated from different sources to meet the desire process requirements of an organization [12]. In most systems, the rate of change to a given master-data entity is fairly low, so update conflicts should be pretty rare. To reduce the chances of update conflicts, concept of best of breed approach has been introduced into system. Best of breed approach defines which system needs to be considered to update the master data, data entities are conflicting with each other. Same customer or person record might have two different data of births in Legacy and CRM. In this case Best of breed approach considers data from CRM system, as it is the place where birth registration information generated from and also it is the most trustable when it comes to DOB. If driving license information is conflicting between legacy and cloud, best of breed approach considers data from Cloud as it is the reliable source of information for this data entity.

V-G) Results- Applying MDM Framework To different applications

The results are recorded by applying the MDM frame work on 3500 sample records of Legacy system, 3018 sample records of CRM and 2374 sample records of Cloud. These applications are citizen centric, where all records related to person or citizen, are considered as master data elements. After applying the MDM framework to the applications, the findings related to the different attributes of the application systems are recorded as specified in following section. Also all these records are corrected by applying the MDM framework.

The percentage of de-duplicate records found in each system are shown in Figure4. After applying the MDM technique, these metrics shows that, percentage of de-duplicate records vary from 64 to 83 in the application systems considered to apply the MDM framework. It also shows that, on an average 27% of duplicate data exists in these systems, which will make data volumes grow large in application system, in turn the maintenance of these systems, go high. By applying MDM technique these duplicates are removed successfully, while loading data into centralized repository.

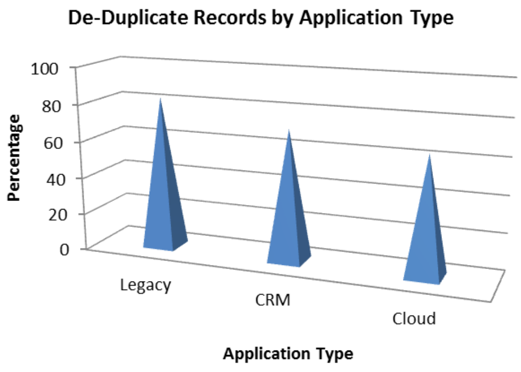

Fig. 4 De-Duplicate records

Results after applying data rules created for E-mail address are shown in Figure5. Data discrepancy for this attribute is 73%, 31%, 56% in Legacy, CRM and Cloud respectively. These discrepancies are identified by removing special characters and spaces from the E-mail address and also by applying best-of-breed approach. This kind of inaccurate data might lead organizations in wrong direction. After applying MDM technique and following the steps specified in implementation part, one can able to identify and correct the E-mail address attribute in MDM repository.

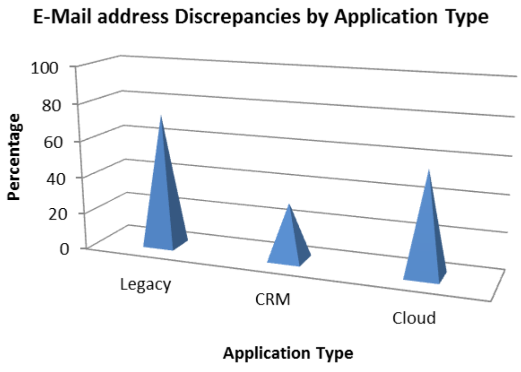

Fig. 5 E-Mail Address discrepancies

Percentage of errors related to first name and last name found in each application systems, are recorded in Figure 6. After applying the MDM technique, these metrics show percentage of error records for attribute first name and last name varies from 11 to 29 in the application systems considered to apply the MDM framework. These discrepancies are identified by, while applying MDM techniques and removed successfully. Maintaining correct data will always help in taking right decisions and following correct strategies.

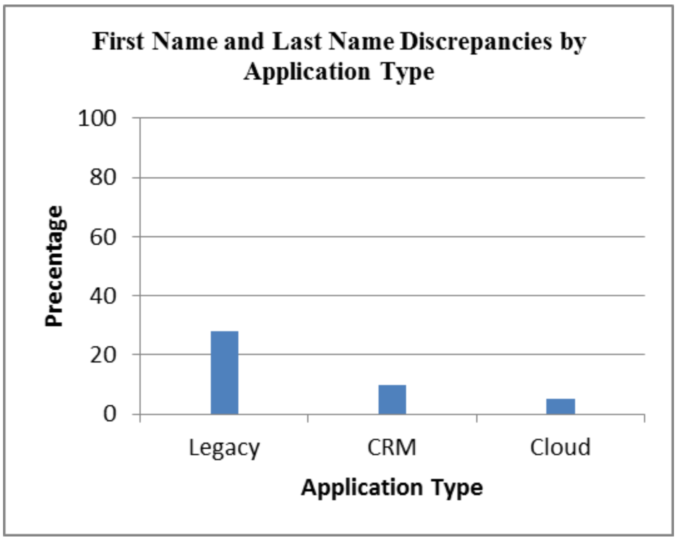

Fig. 6 Percentage of First and Last Name discrepancies

Results after applying data rules created for phone number are recorded in Figure7. Data inaccuracy for this attribute is 29%, 10%, 5% in Legacy, CRM and cloud respectively. These discrepancies are identified after applying data rules. Also these are corrected while loading into MDM repository. This kind of inaccurate data might lead organizations in wrong direction. By applying MDM technique and following the steps specified in implementation part, one can identify and correct these records.

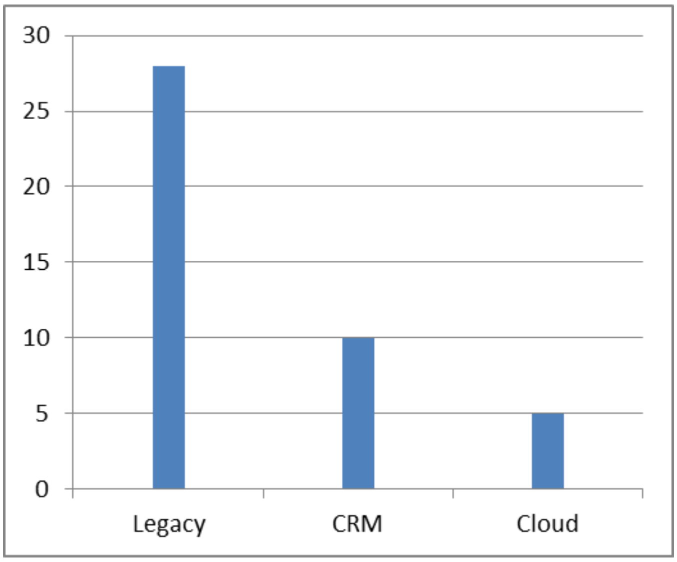

Fig. 7 DOB Discrepancies

Percentage of updates required to apply to phone number attribute in each application systems, are recorded in Figure8. After applying the MDM technique, it is found that, percentage of error records for this attribute varies from 5% to

25% in the application systems considered to apply the MDM framework. These error corrections are related to either old phone number or phone number as null or new phone number is not updated. Data rules framed for this attribute are applied by using MDM technique and these corrected, while loading into MDM repository. Maintaining up-to date data always help organizations to adopt the right choice.

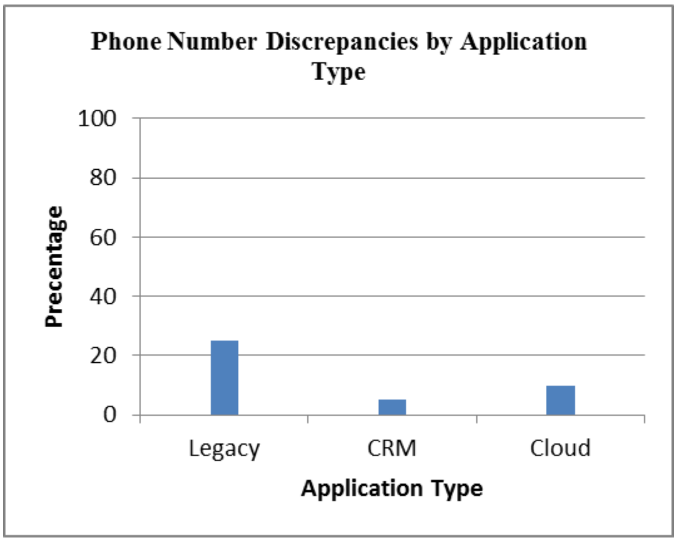

Fig. 8 Phone Number Discrepancies

Below Figure 9 represent the metrics taken by considering all the three Application systems.

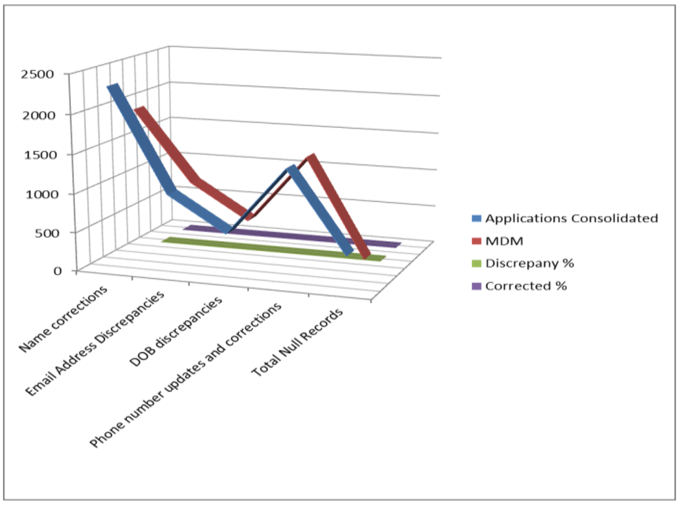

Fig. 9 Discrepancies Metrics ratio for Application Systems

As discussed above, all these discrepancies and inaccuracies are removed successfully, while loading master data elements in to centralized repository. By following MDM framework, it is possible to correct the error records, delete the duplicates and also to keep the central repository up-to date, with single version of truth of the identified attributes. As the master elements are kept in centralized repository, it becomes a painless effort to integrate the existing application systems.

VI. CONCLUSION

In this research work it has been shown that how the MDM framework helps in correcting the error records, deleting inaccuracies, making the data consistent and available to the application in timely manner. By applying this technique to the application systems, it is possible to provide "single version of the truth" for the data collected from various application systems. Now it becomes easy to share data internally and externally, as the information is consistent, accurate and reliable. Instead of maintaining the same data at different places, putting in a centralized repository will save on the infrastructure cost and also the maintenance cost. It is observed from the results that avoiding the duplicates will save lot of space, maintenance and in turn, it saves money to the organizations. In this way MDM framework is incorporated into application systems successfully and also it is observed that these systems can be managed effectively and efficiently by following this approach.

VII. FUTURE ENHANCEMENT

In this work all the data fields have been used to compare the records in order to identify the errors and discrepancies. All application systems are citizen centric system. Images or pictures of the users are also stored in these systems. A system needs to be implemented, where in images of different records across different systems need to be compared to identify errors. This will be very much helpful to the MDM approach to identify same records. MDM framework is applied for Legacy, CRM and Cloud. Likewise in future one can do the same to other application systems like CARD, BHOOMI and Education Departments.

VIII. REFERENCE

[1] IBM White paper "Master Data Management for Big Data"

[2]http://searchdatamanagement.techtarget.com/feature/Enterprise-master-data-management-and-big-data-A-well-matched-pair

[3] http://www.computerweekly.com/feature/Big-Data-and-Master-Data-Management-more-coupled-than-expected

[4] IBM White Paper "Master Data Management and Big Data -- The Keys to Success

[5] Gartner White paper "From MDM to Big Data – From truth to trust",Andrew White.

[6] NESSI White Paper, "Big Data A New world of Opportunities

[7] IDC –Press Release (2012) , IDC Release First Worldwide Big Data Technology and Service Market Forecast, shows Big

Data as Next Essential capability and foundation for the Internet Economy.
[8]Gartner says solving ‘Big Data’ Challenge Involves more than just managing volume of data.
[9] IBM Big Data solutions deliver insight and relevance for digital media
[10].Oracle Big Data Premier-Presentation.
[11]http://wiki.scn.sap.com/wiki/display/EIM/Master+Data+Management+Use+Case