ISSN 2250-3153

Integration and Mapping of IBM Master Data Management to Data Warehouse

Parashiva Murthy B.M.*, Mohan Kera, Bhavani K Eshwar****

*Department of IS & E, SJCE, Mysore.
**MDM, IBM ISL, Bangalore

Abstract **-** This paper presents a solution showing how IBM Master Data Management(MDM-CE) can be integrated and mapped to Data Warehouse. Such integration and mapping benefits MDM-CE and Warehouse users alike as they gain a wider insight into the business and technical information. Raw data from different sources incur a lot of pre-processing before being processed in MDM-CE hence Information Server which can profile and standardize data in the form input to MDM-CE is being integrated with the help of data stage jobs. These jobs basically helps in larger picture of pre-processing and automating data for data load and data export using ETL process of Infosphere Data Stage.

Index Terms **-** Master Data, Data stage, MDM CE, Product Information Management

I. INTRODUCTION

Information is a source of knowledge. But unless it is properly refined and coordinated and available in a proper form for decision making, it is burdensome and not a benefit. Hence information management comes into picture which keeps information organized and distributed in an efficient manner.

Master data is the data that is critical for an enterprise and foundation for key business process and application system and are shared across more than one transactional application. Master Data is not static and has significant number of updates. Usually data at the initial stage is entered manually which is prone to lot of errors hence decreasing the data quality. And this data keeps on changing with time, with modifications and updates from various process and applications. Fixing this poor data quality at its source and managing constant change is very crucial for every enterprise. Global organizations today face a variety of challenges arising from several sources and with time, the nature, scope and size of these challenges change and increase. Master Data Management is a comprehensive strategy to determine and build a single, trusted and authenticated view of a company's information assets and deliver this on demand as a service. MDM helps in reducing the number of errors and also reduces scrap and rework. Hence MDM is the key to the success of business operation.

The main aim here is to ensure quality of master data and seamless integration and leveraging the functionalities of information server by real-time mapping with MDM-CE with the help of data stage jobs which establishes connection and helps in delta/incremental load and data export.

II. IBM MDM-SE OVERVIEW

IBM's MDM-SE is an enterprise-wide MDM platform that provides a virtual view of master data from existing systems. It defines a hub which renders the encapsulation of data domain. It has data model, metadata and actual instance data from sources system. It encompasses vertical data model for the information coming from source. Information which is usually stored horizontally that is all the record information in only one row. With virtualization MDM-SE stores different data into different tables. Hence data access is faster. This is the foundation for IBM MDM-SE software that makes the data model so robust. This kind of mapping demands an abstraction done between the actual physical data location and the hub data model, which is contained in the Data Dictionary.

There is an Inbound Broker that passes source data to the Master Data Engine, which then intelligently derives new data from the incoming record. Compares the information attribute-by-attribute from each record and generates a score for the incoming records and then dynamically links records that meet the pre defined threshold.

The Master Data Engine uses configurable algorithms to create the derivation data, compare member records and produce scores that indicate which records are likely to represent the same entity and the relative strength of the comparison. The key to the Master Data Engine is in how it compares and achieves a finer definition of how two or more records might be related in an enterprise.

III. IBM INFOSPHERE INFORMATION SERVER AND DATASTAGE

IBM InfoSphere Information server enables to understand, cleanse, transform, and deliver context-rich information. It fetches data from different sources, process the data in a required manner and deliver it to target applications. It can work in both batch and real time data flow. The sources may be in different form from each other. The data may be structured or unstructured, coming from either files or databases or any other applications.

Data stage is one of the modules in information server which allows designing job flows. A new stage called MDM stage is

designed to make connection to MDM-CE for data flow. Data Stage job has individual stages linked together that describe the data flow from a data source to a data target. A particular stage generally has at least one data input and/or one data output. Also there can be more than one input/output also. Each stage has properties to define the data processing information such as input file name, processing columns so on. In parallel processing there can be multiple instances of each stage (process) running on the processor. Data stage also leverages the capabilities of parallel processing using multiprocessor platforms and meets the increasing data volumes. Hence maximum throughput can be achieved with such a high performance processing.

IV. MAPPING BENEFITS

Mapping is the process of creating data elementmappings between two distinct data models. Data mapping is used as a first step for a wide variety of data integration tasks including:
(1) Data transformation or data mediation between a data source and a destination.
(2) Identification of data relationships as part of data lineage analysis.
(3) Discovery of hidden sensitive data such as the last four digits social security number hidden in another user id as part of a data masking or de-identification project.
(4) Consolidation of multiple databases into a single data base and identifying redundant columns of data for consolidation or elimination.

For example, a company that would like to transmit and receive purchases and invoices with other companies might use data mapping to create data maps from a company's data to standardized ANSI ASC X12 messages for items such as purchase orders and invoices.

Information server is the one which provides with such functionality were data is cleaned, standardized, transformed in some ways and de duplicated and give a standard form of input to any applications. MDM-SE is the application which performs intelligent processing on the information to provide single trusted view, requires pre processed data in a particular form. Hence by integrating information server ETL functionalities in MDM-SE maximum benefits can be leveraged by reducing the validation burden and even achieving connectivity to MDM-SE and its advantages remotely from any information server. Data can be loaded and exported directly from the data stage job designed.

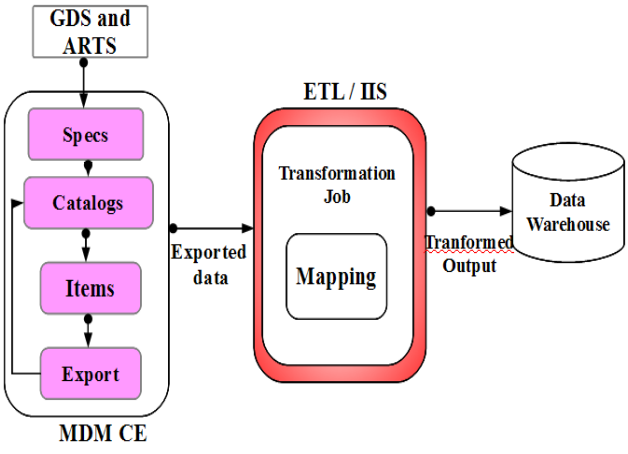
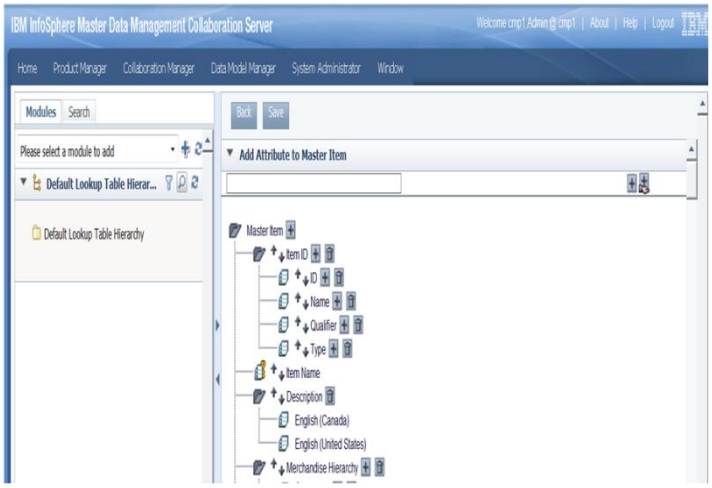


Fig 1- Mapping between MDM-CE and Data Warehouse

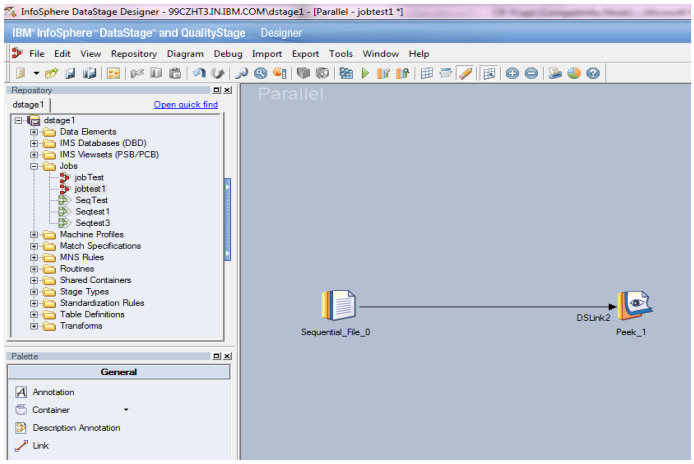
V. PROCESS FLOW

There are three stages in this process - Exporting data using MDM-CE Part, transforming data using Data Stage, Dumping of data to Data Warehouse.

1) *Data Export* - MDM-CE is used for exporting the data. In this stage we create specs, catalogs, items and hierarchies related to the given attributes and finally built the PIM(Product Information Management) model. Finally using export job, The data can be export in any format. For example CSV, Excel, Tab separated XML etc. The exported data is feed as an input to the data stage part.



2) *Data Stage / ETL Stage* - In this stage, the exported data form PIM model can be transformed in to the required model by using transformation features by using ETL as a toolkit. Finally the transformed data is obtained as an output.



3) *Data Warehouse* - The obtained data format from the data stage part has to be dummped on to the data warehouse using netezza. in this process netezza acts as a connector which helps in dumping the data from the source to destination that is data warehouse. Zfinally a kind of mapping happens between MDM-CE to Data warehouse.

VI. CONCLUSION

This paper's aim is to demonstrate the synergy you can achieve by mapping Information server data stage with MDM-SE server. This can be leveraged by a range of users. This is achieved by developing a native export/import utility at the information server side and seamlessly integrating with MDM-SE, and combing the functionality of both products to achieve maximum benefits. This approach could be extended to other IBM MDM versions such as collaborative and advanced editions also. The decisions which are made with inaccurate data, leads to below the optimal performance of any organization. Hence it's very crucial for making intelligent use of the existing resource for maximizing the company benefits.

REFERENCES

[1] IBM SDK Reference for Java and Web Services 2011
[2] IBM Red Books InfoSphere DataStage Data Flow and Job Design, 2008.
[3] Berson, A. and Dubov, L. "Master Data Management and Customer Data Integration for aGlobal Enterprise", McGraw- Hill, New York, 2007.
[4] Lee, Y.W., Pipino, L.L., Funk, J.D. and Wang, R.Y. "Journey to Data Quality", MIT Press, Cambridge, MA 2006.
[5] Schumachar, Scott, "MDM: Realizing the same benefits through different implementations", 2010.

AUTHORS

First Author – Parashiva Murthy B.M, M.Tech In Software Engineering, SJCE, Mysore. Parashivamurthy@gmail.com

Second Author – Mohan Kera, IBM ISL.

Third Author – Bhavani K Eshwar, IBM ISL.