# A Unified and Scalable Data Migration Service for the Cloud Environments

B. Gopi Krishna, E. Vengal Reddy, K. Jagadamba, Srikumar Krishnamurthy, P. Radha Krishna

Software Engineering and Technology Labs,
Infosys Technologies Limited, India
{ Gopi_Balasa, EggonuVengal_Reddy, Jagadamba_Krovvidi, Srikumar_K, RadhaKrishna_P}@infosys.com

## Abstract

Data migration is one of the key operations in many enterprise data integration projects. While there are standard and well established tools for data migration, the enterprises recent move towards cloud environments present increasing challenges with respect to scalability, availability and data model heterogeneities. In this demo paper, we present a scalable data migration service that employs automated schema matching techniques to handle the schema disparities in cloud and enterprise data models. The proposed system also uses elastic grid infrastructure for on-demand data access, rules driven transformation and on-the-fly integration of distributed data sources for migration. We bring out key research challenges in building such a system and present viable solutions for addressing the same. The demo is likely to be useful for the academicians in getting insights on newer challenges in this field and the practitioners can benefit from understanding new offerings in this space

## 1. Introduction

Cloud computing is an emerging computing paradigm that was innovated to deploy cost effective solutions over Internet. Companies such as Google, IBM, Amazon, Yahoo and Intel have already started providing computing infrastructures for variety of data processing and data storage applications on a pay-per-use basis. Besides, organizations have also started porting their data and applications to cloud in order to reap the benefits of this new paradigm. This necessitates a well-defined data migration as a service (DMaaS) over the cloud environments. On the other hand, cloud utility is facing problem due to lack of trust and standards around security and privacy. Some of these can be overcome by building a corporate cloud. Given a cloud with certain amount of data, we need solutions to migrate this data to another cloud. Thus, various scenarios can be visualized for migrating data with-in and around cloud: (i) enterprise to cloud data migration and vice versa and (ii) inter-cloud data migration.

Data Migration into and from cloud environment is a challenging task. Migrating data from legacy applications to cloud needs significant efforts and it is one of the major factors in deciding to use cloud computing [1]. The major cloud service providers currently providing services such as application services, storage services, compute services and data services [2]. Our work focuses on data services in general, and data migration in particular. Amazon SimpeDB, Microsoft SQL server data services and Google's data store provides services for data storage and access for their cloud data. However, there are several issues with regard to data migration from or into the cloud environment. The major issues in representing and accessing from cloud databases are: (i) Cloud follows column oriented databases and (ii) do not have relational concepts such as foreign keys (for joining) and indexing. A detailed comparative characteristic of data migration in normal and cloud environments are presented in Table 1.

Data Migration has to deal with the dissimilar technologies, multiple traditions, architectures and practices. The business reason for migration will decide the complexity of data migration. The proposed system enables Data Migration as a service with appropriate migration accelerators, business solutions and connectors. Figure 1 shows various scenarios for data migration service in a cloud environment. Our research prototype aims to address the challenges in data migration for cloud environments. The presented prototype system is unified and scalable in terms of bulk data transfer and distributed query processing capabilities, besides offering relational schema mapping from/to enterprise and cloud environments.

*15th International Conference on Management of Data*
*COMAD 2009, Mysore, India, December 9–12, 2009*

Table 1. Characteristics of Data Migration System

| Parameter | Traditional Data Migration | Cloud Data Migration |
|---|---|---|
| **Data Access and Storage** | | |
| Accessibility / Connectivity | Target Databases can be accessed using Open Database Connectivity (ODBC / OLEDB Drivers). | Cloud Data Storage can be accessed through specialized services (ex. Windows Azure Account) as the system on which data gets hosted is not a user choice. |
| Querying Or Information Retrieval | Relational Targets have a strong querying through standard SQL. Information also can be retrieved in a meaningful way for presenting it to the external world. | Yet to evolve robust querying mechanism for Cloud Storage. |
| Information Storage | Relational Systems does provide proper source where information was stored physically. | Cloud Architecture hides the physical storage details from end-user. |
| **Functional Aspects** | | |
| Relations | Target Databases are relational in nature, hence need to adhere to the defined relational constraints | Cloud Storage does not support relations. |
| Table/Entity Structure | Target Databases have Fixed Table / Entity structure, which cannot be changed during the course of migration | Cloud Storage accommodates dynamic changes to the entity structure. Data with different source entity structures can be accommodated in a single cloud entity. |
| Data Model | Data Model will be designed externally | Cloud does not support any specific data models. It is flexible and it can take any source model. |
| Metadata | Metadata can be extracted using data dictionaries. | Cloud does not have any Metadata extraction mechanism. |
| Mapping | Source entity attributes can be mapped to the corresponding target entity attributes. | As the storage is based on objects, the mapping process is even more complex |
| **Non-functional Aspects** | | |
| Performance | Data Loading time will be as per the ETL Server capabilities. | Performance could be slow as data loading happens from on-premises to public cloud over internet by crossing the firewalls. |
| Security | Data Movement can be done in a secure environment | Cloud Platforms yet to address the security concerns. One solution is to have the corporate (private) clouds, provided the budget permits. |

## 2. System Description

We present a scalable and unified data migration service for cloud environments. Figure 2 shows the high-level functional diagram of our prototype system (iMigrate++). The key components of this architecture include:

- A distributed query processing engine that processes queries using a scale-out grid infrastructure,
- Business rules modeler and rules engine for capturing custom business rules for migration,
- An automatic schema mapping and translation engine for schema migration, and
- A data migration engine that provides data discovery, exploration and transformation services

The research prototype enables mapping a relational or xml database schema to corresponding cloud data entities. These mapping constructors facilitate data transfer services from distributed relational databases to cloud.

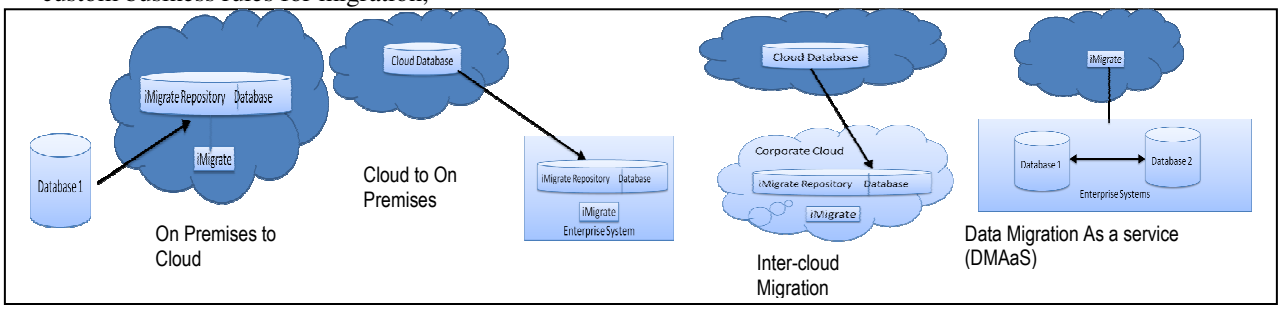


Fig 1. Scenarios for data migration over cloud

## 3. System Novelties

Some of the unique aspects of the proposed system are:

1. Auto Script generation for Mapping (Relational and XML) Data Elements to Cloud Entities
2. Microsoft Business connectors for interoperability between Commercial Databases
   The system supports Microsoft business connectors (AXAPTA, CRM, Azure Cloud, Infosys Simple DB) for migrating data to MS ERP / CRM Platforms.
3. On-Demand Bulk Data Transfer
   The System can be hosted as an application on Microsoft Cloud Azure Platform and can be used to enable bulk data transfer on cloud. This solution also uses best data loading utilities (such as Oracle External Tables, SQL Loader, Data Pump, SQL Server Bulk Copying, and DB2 Bulk Load Utilities) appropriately so that data loading will be at par with the database server load capabilities.
4. Business Rule Modeling and Processing
   Business Rules Engine processes and transforms output that can be migrated to the target systems.
5. Patent pending technology to map heterogeneous schemas semi-automatically
6. Identify complex mappings using rich source of instance information, past matches, business rules and business vocabulary
7. Advanced distributed query processing engine that utilizes elastic grid infrastructure to provide scalable and on-demand data aggregation from multiple heterogeneous enterprise systems.

## 4. System Features

The main features of the system can be broadly classified into four: are:

- Data Access ( Data extractor, Metadata extractor and validator, Cloud connector)
- Data Processing ( Data quality, Data validation)
- Data Migration and Analysis (Design accelerator - Schema mapper, Mapping builder, Distributed query processing)
- Data management (Job monitor)

Below we describe briefly the functions and features of the system:

**Data Extractor** unloads the data from source databases into delimited text files. iMigrate++ Repository Staging Tables will be populated based on the extracted files.

**Metadata Extractor and Validator** component extracts metadata such as table definitions, column definitions, constraints, data rules from source and target databases. The validator component finds and generates a report with the discrepancies between source and target metadata definition such as Datatype, Data Size and Constraint mismatches.

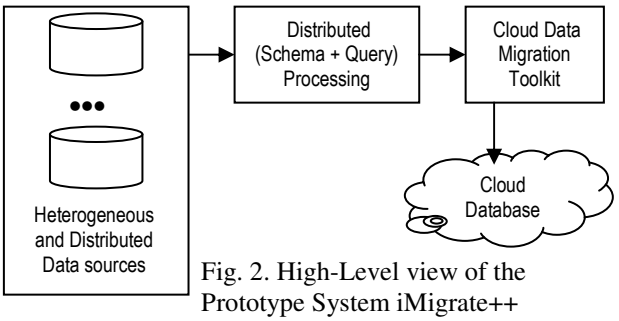


Fig. 2. High-Level view of the Prototype System iMigrate++

**Cloud Connector** uses Microsoft Windows Azure Table Storage. Windows Azure Storage allows application developers to store their data in the cloud. The application can access its data from anywhere at any time, store any amount of data and for any length of time, and be confident that the data is durable and will not be lost. Windows Azure Storage provides a rich set of data abstractions:

**Data Quality Accelerator:**

Data Quality Assessment and Enrichment features that are supported by the system include:

- Pattern and suspicious data analysis and fixing.
- Data Discovery.
- Redundant data analysis and removal.
- Data Consolidation and Comparison.
- Relationship Analysis.

**Data Validation Accelerators:**

**Job Statistics component p**rovides Load Statistics such as Job Start Time, Finish Time, Number of Rows Extracted / Inserted / Updated / Deleted / Rejected, Error Code / Description, Job Status.

**Test Case Generation module** has testing module where Data Migration Developer can create ample test cases with testing module and can generate test case comparison report as per the user defined criterion.

**Metadata Validation Reports module** provides a metadata validation report by comparing source schema Vs Target Schema at the mapping design level, so that the discrepancies can be corrected before the actual data load, which saves lot of job loading time as the production load window is critical.

**Design Accelerator** has many sub-components like

*Schema Mapper*: Generates mapping automatically between the source and target tables using schema element, structure, instance data, business vocabulary and business rules.

Design Import and Export: Data in other formats (ex. MS Excel) can be easily imported and exported to iMigrate++ Tool.

**Script Generator** component generates target database specific script based on the registered metadata and mapping logic between the source and target systems. It

also generates the code required for capturing the runtime load statistics and error logging.

**Job Monitor** provides an interface for executing the iMigrate++ generated scripts. It also facilitates the end user to re-set the job status in order to initiate the load process again. It also displays the recent job-runtime statistics like job start time, end time, number of rows extracted, inserted, rejected and error messages if any.

Windows Azure Blob – provides storage for large data items.

Windows Azure Table – provides structured storage for maintaining service state.

Windows Azure Queue – provides asynchronous work dispatch to enable service communication.

*Mapping Builder* provides an interface to map the source / target tables and columns. It also has a transformation builder which has basic data migration transformations such as expression builder, row key generator, lookup and function invoker.

## 5. System Use Case

In order to illustrate the capabilities of the system, we will present the case of automobile company. ABC Company, a leading automobile organization, is planning to transform their existing systems into a single consolidated view to cater the current business model needs. Most of the ABC Business applications and portals are developed using ASP.net, VB.net and they are using several databases. The company expects a SaaS business model to cut down the total cost of ownership (TCO) and does not want to invest heavily. The proposed solution should address their customers to subscribe to the company services and the solution should address company to publish their services. Also, it should facilitate the key business managers to view and track the business at their respective regions. ABC Company wanted to outsource their infrastructure needs and maintenance activities and ready to pay as much as they use. Here are high level business goals summarized:

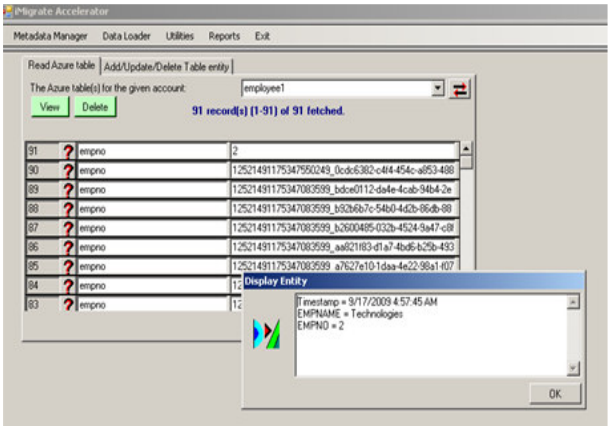


Fig. 3. A snapshot of data migration to Microsoft Azure Table Storage

- Needed an integrated business system for the single version of truth
- Wanted to cut down the cost as much as possible
- Facilitate for publishing the company services such as Product Availability, new launches, warranty, features, other services
- Facilitate customer requests for new product enquiries, online purchases, service.
- Needed a SaaS Model
- Wanted to outsource the application infrastructure set-up and the maintenance

A proof-of-concept (POC) has been developed using Microsoft Azure Cloud platform to migrate their existing data in Oracle and SQL Server databases into a cloud. Figure 3 shows the key-value pairs after migration of data into Azure platform. The major activities of this POC are: (a) Queries disparate data sources and extracts relational metadata, (b) Registers the extracted metadata in iMigrate++ repository model, (c) Facilitates Data Migration Developer to prepare the source mapping query, (d) Migrates integrated data from disparate sources to temporary Staging Area for Assessment, (e) Executes Data Quality Assessment Queries, (f) Transforms the extracted data, cleanses as per quality enrichment, (g) Connects to Microsoft Azure Windows Live Account, (h) Creates Azure Entity on Cloud Platform, (i) Creates a collection of Azure table entities with data from the database tables and (j) Validates The Data Migration and Presents Load Statistics.

The proposed solution facilitates faster migration service and quality enrichment of the data, besides availing normal cloud benefits such as cost and elastic demand of resources.

## 6. Conclusions

We presented a unified and scalable data migration system for cloud environments. The proposed system aims to overcome some of the limitations of traditional migration systems. It provides a Scalable data federation engine to migrate enterprise data to cloud environments. In addition, it provides automated schema mapping engine that can identify complex mappings using advanced meta-heuristic algorithms. The proposed system is likely to be useful for both academicians as well as practitioners.

**Reference**

1. Michael, et. al, About the Clouds: A Berkeley View of Cloud Computing, Technical Report No. UCB/EECS-2009-28, Feb. 2009.
2. Rakesh Agarwal, et al., The Claremont Report on Database Research, SIGMOD Record, Vol. 37, No. 3, pp. 9-19, 2008.