Statistical Comparison of Architecture Driven Modernization with other Cloud Migration Frameworks and Formation of Clusters

Mubeen Aslam*1, Lukman AB Rahim*2, Manzoor Hashmani*3, Junzo Watada*

*High Performance Cloud Computing Centre, Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia
E-mail: 1mubeen.aslam_g03587@utp.edu.my; 2lukmanrahim@utp.edu.my; 3manzoor.hashmani@utp.edu.my

*Department of Computer and Information Sciences, Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia
E-mail: junzo.watada@utp.edu.my

*Department of Computer Science, National Textile University, Faisalabad, Pakistan
E-mail: shim2saqjid@gmail.com

Abstract—Corporations are migrating their legacy software systems towards the cloud environment for amelioration, to avail benefits of the cloud. Long term success of modernizing a legacy software depends on the characteristics of the chosen cloud migration approach. Organizations must think over how strategically imperative is the chosen cloud migration framework to their business? Thus, the Object Management Group (OMG) has defined standards for the modernization process based on Architecture Driven Modernization (ADM) framework. ADM serves as a vehicle for facilitating the arrangement of information technology with business stratagem and its architecture. Until now, it seems that there is no systematic mapping among ADM and other cloud migration frameworks, highlighting the demanding features. This research aims to give an in-depth study of similar cloud migration frameworks. Thus, the researchers introduced the clusters containing cloud migration frameworks having similar features to ADM. This systematic mapping can be seen as a valuable asset for those who are interested in choosing the best migration framework from the pool of cloud modernization frameworks, according to their legacy software requirements. The clustering technique is used to appraise and compare ADM with some of the other cloud migration frameworks for highlighting the similarities and key differences. The quality of clusters is evaluated by the Rand index and Silhouette measurements. The study distills the record and yields a sound and healthy catalog for essential events and concerns that are communal in cloud migration frameworks. This research offers the one-stop-shop convenience that the industry desperately desires.

Keywords—cloud migration frameworks; Architecture Driven Modernization (ADM); statistical analysis; clustering techniques.

I. INTRODUCTION

Cloud computing assures the on-demand scalability of computer technology. Legacy software systems can be seriously challenging for organizations because the legacy system cannot be discarded since they store a lot of valuable business information, and on the other hand, these legacy applications cannot be maintained economically. To save all the investments done on outdated systems, organizations are enthusiastic about shifting them to the cloud [1]. Cloud migration is not easy for everyone. Cloud environments are generally reliable and highly available. These are not only the considerations while migrating an application. For migration of a system, there are many factors to consider, from benefits and risks [2], cloud service model, type of language used [3], type of migration done, modernizing product, operational cost reduction, the required requirements. This erudition enforced the researchers to dig into the research to figure out the best and easiest way to find the desirable cloud migration framework. Organizations choose a migration strategy [4] for modernizing their legacy systems, based on their requirements. Significance and usefulness of the ADM framework given in literature have diverted the researchers’ attention in finding out the similarities and key differences among ADM and some of the other commonly used cloud migration frameworks. A Clustering method named Partitioning [5], an unsupervised learning way is used to discover a new set of categories of cloud migration frameworks. It is done by applying an in-depth cluster analysis and statistical methods.

Formally and conventionally, the clustering structure can be represented as a set C of subsets C1, C2, , , Cn, such as C1 ∩ C2 ∩ C3, , ∩ Cn = θ. C (C1,C2, , ,Cn) means that each subset has a specific proportion of similar key features to ADM. Instances of each subset clusters are similar to each
other and are different from the instances of the other subset clusters. Higher Similarity Value is the fundamental feature of a clustering process [2]. Jaccard [6], K-Means [7], Euclidean, and Louvain statistical measures are used to find the similarity among ADM and other cloud migration frameworks. These measurements produced five clusters, which are evaluated by the Rand Index and Silhouette evaluation methods. Satisfactory results validate the quality of the clusters produced. The objective of this research is to divide an unorganized and considerable pool of cloud migration frameworks into smaller pieces/clusters to reduce the overall processing and investigation time.

II. MATERIALS AND METHOD

A. Cloud Migration Frameworks

For a successful migration of an application from one platform to another, the researchers have to follow the steps of a proper framework [8]. Therefore, the choice of a migration framework matters a lot. Each migration framework differentiates from one another having different types of features like the type of; the language used to build up the framework, migration strategy, transformations made, and the technologies used.

1) Architecture Driven Modernization: Huge data, code, documentation, outdated technology, etc are the main challenges faced by the migration frameworks. ADM mostly resolves all these issues as it focuses on modernization activities based on architectural models rather than code artifacts [9]. According to [10] to architecture is the fundamental organization of a system embodied in its components, their relationships to each other and to the environment and the principles guiding its design and evolution. The concept which helps us to understand and make the system reusable is the Architecture Driven Modernization (ADM), through which the researchers can transform the architecture of the existing legacy system to the new one, as shown in Fig 1.

ADM can be used for software improvement, interoperability, reuse, modifications, restructuring, migration, translation, integration, and service-oriented architecture. The transformations (T1, T2, T3) performed in the ADM framework for cloud, goes from Code – Model – Code (C2C), as shown in Fig. 2.

In it, T2 is the detail description of T1 and T3 is the detail description of T2 shows the complete transformations performed in the ADM migration framework. ADM is a combination of ADM and MDA.

2) Other Cloud Migration Frameworks: Cloud migration frameworks chosen for this research are the most used by the cloud practitioners in literature. Sabiri et al. agree to have a classification of migration issues i.e., decision making support [11], migration methods and development tools. Agile model-driven modernization (Agile MDA) methodology is approved by [12][13] (RMA7 in Table I) for the SaaS platform by using REMICS methodology by metamodel and UML under migration complete, also acknowledged by [15], [19]. They proposed that REMICS can be seen as a Model-Driven Black-Box modernization approach with flavors of white-box migration in particular for legacy system transformation [15]. All three technologies combined (REMICs + MDA + Agile) [18] (RMA8 in Table I) are broadly researched and applied. By analyzing all the above REMICS research work, it is concluded that REMICS has borrowed several concepts from OMG, like ADM focusing on recuperating the legacy system artifacts in the early stage of transformation [20]. It is challenging to move applications from one platform to the other [21] i.e., from IaaS to PaaS or SaaS. MODA Clouds [11] support system creators and operators in using multiple clouds for the same system and in migrating a portion of the system from cloud to cloud. Using the REMICS [12], [15],
[18] approach with MDD supports the migration of existing software to the cloud. MODA Clouds used Cloud ML DSML under the migration strategy of Iterative on the cloud. It goes from CIM (independent cloud model) CPIM (cloud provider independent model) CPSM (cloud provider-specific model) which is a Model-Model architecture, rather than ADM. Table 1 gives the brief description of frameworks and their Language (Metamodel, UML), Transformation (M2M, C2C, M2C), Migration type/strategy (Holistic, standard format, component format etc.), Platform (SaaS, PaaS, IaaS) and Modernizing product (legacy, new).

B. Clustering

Larger the similarity or homogeneity within a group and smaller the difference between the groups, more concrete are the clusters [22]. How well the groups are related to each other can be measured by distance measurement [23]. Cluster analysis [24] can be a powerful data mining tool for any organization that needs to distinguish groups like fraudulent claims, credit scoring, and items in a grocery store. There are two types of measures i.e., Similarity Measure and Distance Measure. Both types have multiple techniques like Hierarchical, partitioning, density-based, grid-based and soft-computing [4]. The sub-techniques used for this research are: Jaccard similarity, K-means, Euclidean Distance and Louvain Clustering.

C. Research Methodology

This research utilizes both qualitative and quantitative methods. This section presents the methodology applied to this research. This research is based on three main sections i.e., Identification, Similarity, and Dissimilarity Measurement and Verification. Each section has different phases shown in Fig. 3.

1) Identification: In this first section, the most demanded and the most used cloud migration frameworks were identified. The next step was to extricate the features of each framework.
   - Cloud Migration Frameworks: Cloud Migration Frameworks were chosen based on the current market usage and demand, (as mentioned in Table I) like REMICS, CloudMIG, and ARTIST.
   - Feature Detection: The most used features (as shown in Table I) required for the migration of an application on the cloud, were detected, like language, transformation type, and migration type.

2) Similarity and Dissimilarity Measurement: This is the phase where the similarity between migration frameworks is investigated. The researchers used the manual calculations and a simulation tool named Orange for the clustering algorithms and the verification process. The details of the statistical measurements are as follows:
   - Jaccard Similarity Clustering: It is a statistical measurement that measures similarities between sets.

---

### Table 1: Features of Cloud Migration Frameworks

<table>
<thead>
<tr>
<th>Frameworks</th>
<th>ADM/ MDE/ MDA</th>
<th>Language</th>
<th>Migration Type</th>
<th>M2M/ M2C/ C2C</th>
<th>IaaS/ PaaS/ SaaS/ IaaS</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODA CLOUDS</td>
<td>MDD</td>
<td>MODAClouds ML DSML</td>
<td>Iterative Migration</td>
<td>M2M</td>
<td>IPS</td>
<td>[11]</td>
</tr>
<tr>
<td>CloudMIG</td>
<td>ADM</td>
<td>UML</td>
<td>Revise</td>
<td>M2M</td>
<td>SaaS</td>
<td>[14]</td>
</tr>
<tr>
<td>REMICS</td>
<td>ADM</td>
<td>Metamodels + UML Profile</td>
<td>Replace / Wrapping</td>
<td>M2M</td>
<td>SaaS</td>
<td>[15]</td>
</tr>
<tr>
<td>ADM</td>
<td>ADM</td>
<td>Metamodels</td>
<td>Holistic / Reengineering</td>
<td>M2M</td>
<td>IPS</td>
<td>[9]</td>
</tr>
<tr>
<td>Relative analysis</td>
<td>MDE</td>
<td>Metamodel</td>
<td>Holistic</td>
<td>M2M</td>
<td>IPS</td>
<td>[16]</td>
</tr>
<tr>
<td>ARTIST</td>
<td>MDE</td>
<td>CAML- UML</td>
<td>Cloudify</td>
<td>C2C</td>
<td>SaaS</td>
<td>[17]</td>
</tr>
<tr>
<td>RMA7</td>
<td>MDA</td>
<td>Metamodel + UML</td>
<td>Complete</td>
<td>M2M</td>
<td>SaaS</td>
<td>[12]</td>
</tr>
<tr>
<td>RMA8</td>
<td>MDA</td>
<td>UML</td>
<td>Component Format</td>
<td>M2M</td>
<td>SaaS</td>
<td>[18]</td>
</tr>
</tbody>
</table>

---

![Fig.3: Research Methodology](image-url)
It is defined as the size of the intersection divided by the size of the union of two sets. It is also named as Jaccard Coefficient.

\[
\text{Jaccard Coefficient} = \frac{|A \cap B|}{|A \cup B|} \quad (1)
\]

- **K-Means Clustering**: K-means algorithm is one of the top recognized, standard and simplest clustering algorithms and is commonly used in multiple areas. In it, k - centroid is defined for each cluster [25]. The goal of this method is to create a specific number of groups in the data. The cluster creation is based on feature similarity.

- **Euclidean Clustering**: Euclidean distance is a simple distance between two points. It is usually used for clustering tech. The default distance measured with the K-means algorithm is also the Euclidean distance. Equation 2 shows the Euclidean distance function which is the root of squared differences between the coordinates of a pair of objects [26].

\[
\text{Dist}_XY = \max k | X_{ik} - X_{jk} | \quad (2)
\]

- **Louvain Clustering**: It is a well-known fast and efficient modularity-based graph clustering algorithm with near-linear runtime in sparse graphs. Due to its fast and efficient results, it is used in some applications such as social media [27] [28]. Equation 3 [29] shows the modularity to measure;

\[
M = \sum_{c} \left( \frac{\sum_{n} \frac{e_{cc}}{2m} - \frac{(\sum_{e} e_{cc})^2}{4m^2}}{2m} \right) \quad (3)
\]

3) **Verification**: The final stage was to verify all the results and the clusters. The two well-known evaluation methods were used i.e.

- **Rand Index Verification**: Establishment of the accurate cluster is a challenging task and it is also imperative to test the validity and accuracy of the clusters. Quality to provide the information can be done in different ways like merging sig sigma approaches in the STOPE view of an organization [30]. This helps in forming a group/cluster of similar characteristics. Clusters should be tested in a way to find out whether they show the maximum similarity among the objects in the same cluster and minimum similarity among those in other clusters. Rand Index (RAND) is the most straightforward and often used criterion to verify the level of similarity of the clusters [31]. The RAND is defined in equation 3.

\[
\text{RAND} = \frac{2TP+TN}{TP+FP+FN+TN} \quad (4)
\]

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. The value of the rand index lies between 0 and 1. Where there is an agreement between two objects, the RAND value is 1. This research found 1 for all the clusters which verify the quality of the cluster.

- **Silhouette Plot Verification**: Silhouette Index (SI) is a metric that compares the distance of one point to another point in the cluster with the distance between two points of the closest cluster. The value of Silhouette varies in the range [-1, 1]. The result shows near to 1 is the best result [32], [33]. It is computed as in equation 5.

\[
\text{SI} = \frac{1}{N} \sum_{i=1}^{N} \frac{(b_i-a_i)}{\max(a_i,b_i)} \quad (5)
\]

Where;

- \( x_i \) = elements in a cluster.
- \( a_i \) = average distance from one element to all the other elements in its cluster,
- \( b_i \) = minimum of the average distance of an element to all the elements of the other cluster.

### III. RESULTS AND DISCUSSION

The researchers tried to group the most similar cloud migration frameworks by applying different clustering algorithms. For this, Jaccard Coefficient was applied on the data set in Table I. Table II gives the aliases and the values assigned to the features selected from the cloud migration framework of Table I.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Sub attributes</th>
<th>Aliases</th>
<th>Values assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Metamodel</td>
<td>M</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Unified Modeling Language</td>
<td>U</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Both M and U</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Transformation</td>
<td>Model to Model</td>
<td>M2M</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Code to Code</td>
<td>C2C</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Model to Code</td>
<td>M2C</td>
<td>5</td>
</tr>
<tr>
<td>Migration Type</td>
<td>Holistic</td>
<td>H</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Standard Format</td>
<td>SF</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Component Format</td>
<td>CF</td>
<td>8</td>
</tr>
<tr>
<td>Platform</td>
<td>Software as a Service</td>
<td>SaaS</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Platform as a Service</td>
<td>PaaS</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Infrastructure as a Service</td>
<td>IaaS</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Software, Platform, and Infrastructure</td>
<td>SPI</td>
<td>12</td>
</tr>
<tr>
<td>Modernizing Product</td>
<td>Legacy</td>
<td>L</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>N</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Both Legacy and New</td>
<td>B</td>
<td>15</td>
</tr>
</tbody>
</table>

Table III presents the calculations performed by the Jaccard coefficient (Eq. 1).
The percentage in Table III shows the range from highly similar framework to the least similar framework. It illustrates that the Relative Analysis Migration Framework has the most similar features to ADM. Figure 4 below presents the clustering algorithms used in the Orange tool. The file is the dataset file. K-means, Euclidean and Louvain clustering are the algorithms applied on the data file. Scatter Plot and Distance Matrix show the calculated results. Silhouette Plot presents the validation results. Figure 5 displays the clusters based on the Jaccard and K-mean measurements done on the data file. Different colors are assigned to the clusters, by the Orange tool.

Similarly, Figure 6 represents the Louvain Clustering displaying colorful clusters. Figure 7 represents the Euclidean Clustering result by displaying the Distance Matrix. It is a table that shows the distance between pairs of objects. This table has a leading diagonal, which is always 0 and is shown as blank cells. This 0 is the object’s distance from itself and is ignored. The upper triangular part of the matrix is just a mirror of the lower triangular and is ignored as well. This matrix displays 0.000 distance between CloudMIG and RMA8, and MODACLouds and Artiste. The rest of the frameworks have small distances.
Fig. 5: Jaccard and K-Means Clustering

Fig. 6: Louvain Clustering
After having all the results from the clustering algorithms, the clusters formed are: 1=Relative Analysis, 2=REMICS, 3=RMA7, 4=CloudMIG and RMA8, and 5=MODA-CLOUDS and ARTIST, presented in Figure 8.

Number 1 is the most similar cluster to ADM whereas number 5th is the least similar migration framework to ADM. This ranking is based on the end results produced by the clustering algorithms. These clusters can help the industry to have increased productivity, more rapid innovation and can expand their boundaries. They can match their requirements with these clusters and can choose the best cloud migration framework. This reduces lots of time.

Verification for the quality of clusters was made by Rand Index (Eq. 4) and Silhouette Plotting by Orange. Figures 9 and 10 show the results of Silhouette plot for Jaccard and K-Means and Louvain clustering. Figure 11 shows the calculated Rand Index. As explained in Section II, the best results for Silhouette and Rand Index is 1. Figures 9,10 and 11 show satisfactory results and thus verifies the quality of clusters.

IV. CONCLUSION

In this paper, the researchers present the qualitative and quantitative study to identify the similarity among cloud migration frameworks with ADM. ADM is taken as a standard cloud migration framework. The researchers first identified the most used cloud migration frameworks. Next, the features of each migration framework were inspected.
Different clustering algorithms like Jaccard, K-Means, Euclidean, and Louvain were employed on the features. The researchers compared the cloud migration frameworks. Statistical calculations and a simulation tool named Orange were used to produce the results. The similarity index produced clusters containing similar cloud migration frameworks. These clusters are ranked from 1-5, presenting from a higher similarity level to the lowest. The researchers can conclude that there was a need for these clusters to improve productivity and reduce the operational cost. So that cloud consumers do not worry about choosing the migration framework and can concentrate on other issues like developing and monitoring.

The researchers evaluated the quality of clusters by applying Rand Index and Silhouette verification methods. The satisfactory quality results evidenced the reliability of the clusters. This paper enables both academia and practitioners in the cloud computing community, to get a predominant view of the migration strategy for legacy application towards the cloud and for investigating new research areas.

REFERENCES


