National Sport Institute Case: Automated Data Migration Using Talend Open Studio with 'Trickle Approach'

Suzaimah Ramli[1(✉)], Izzuddin Redzuan[1,2], Iqbal Hakim Mamood[2], Norulzahrah Mohd Zainudin[1], Nor Asiakin Hasbullah[1], and Muslihah Wook[1]

[1] National Defence University of Malaysia, Kuala Lumpur, Malaysia
suzaimah@upnm.edu.my
[2] National Sport Institute, Kuala Lumpur, Malaysia
iqbal@isn.gov.my

Abstract. In the current era of globalization, old systems are replaced or improved daily. Before the transition happens, IT experts need to carefully plan the best course of action to move the massive amount of data from old to new systems. If there are flaws unnoticed, the data may be lost during the migration process where it will feed unreliable data to the end-users. Recently, National Sports Institute had launched a new system the National Coaching Academy System where it was outsourced and developed by a vendor. Unfortunately, the system cannot be used immediately because there are no data from the old system in its database. This kind of problem occurs when data migration is not considered during the development phase of a new system. Because of this, the staff need to manually navigate between the two old and new systems to obtain accurate data of their coaches. Thus, in this paper, we will discuss the process that needs to be analyzed and what are the best strategies to solve the problem. The result should be able to increase the efficiency and make it easier for the staff when working with the coach's data.

Keywords: Automation · Data migration · ETL process · Talend open studio

1 Introduction

Data migration is the process of transferring data between storage systems, data formats, or computer systems. Data migration projects are undertaken for a number of reasons, including changing or upgrading servers or storage devices, transferring data to third-party cloud providers, merging websites, maintaining infrastructure, migrating applications or databases, updating software, corporate mergers, or relocating data centers [1]. A project that involves data migration has been quiet challenging to most IT experts where they must time the project so there is minimal impact to an organization. Depends on the budget, not only they have to keep an eye on the cost but also the time taken for the project to finish. IT experts need to allocate most of their time to maintain

H. Badioze Zaman et al. (Eds.): IVIC 2021, LNCS 13051, pp. 214–223, 2021.
https://doi.org/10.1007/978-3-030-90235-3_19

data integrity to avoid problem occurs during the migration where it will risk of a prolong disruption and downtime to an active organization.

As part of the extract/transform/load (ETL) process, each data migration includes at least the transformation and loading steps. This means that the extracted data must go through a number of preparatory functions before it can be loaded into the target locale. Companies migrate data for several reasons. You may need to upgrade the entire system, upgrade the database, build a new data warehouse, or combine new data from acquisitions or other sources. Data migration is also important when implementing other systems in addition to existing applications [2]. The data must be relevant for the purpose of the new system. This often requires data validation, correcting database problems, changing data formats, or combining values.

Talend Open Studio is an open architecture for data integration, data profiling, big data, cloud integration and data migration. It is a GUI environment that offers more than 1000 ready-made connectors. This makes it easy to perform operations such as changing files, loading data, moving files, and renaming files. This allows each component to define complex processes. Integration tasks are created by configured Talend components, not coded. In addition, tasks from the development environment or as stand-alone scripts can be executed [2]. Talend offers components that include database administration, auditing, and monitoring. This component helps with the administration of user accounts, permissions, and project permissions. The audit database helps evaluate various aspects of the workplace to develop the ideal process-oriented decision support system.

2 Literature Review

2.1 Type of Data Migration

There are a few types of migration that can be done depends on the organization needs such as the storage migration moves data off existing arrays into more modern ones that enable other systems to access it. It offers significantly faster performance and more cost-effective scaling while enabling expected data management features such as cloning, snapshots, and backup and disaster recovery. Next is cloud migration where it moves data, application, or other business elements from either an on-premises data center to a cloud or from one cloud to another. In many cases, it also entails a storage migration. Application migration moves an application program from one environment to another. It may include moving the entire application from an on-premises IT center to a cloud, moving between clouds, or simply moving the application's underlying data to a new form of the application hosted by a software provider [3]. The importance of choosing the correct strategy is crucial in a data migration project due to the compatibility of the system especially when the system uses a different architecture and formats.

2.2 Data Migration Strategies

Organizations need to consider which data migration strategy will best suit their needs. Depending on the project requirements and the available editing windows, you can choose from several strategies. There are two main types of migration: "Big Bang" and

"trickle". When migrating Big Bang data, all data is migrated in one operation. While this may take some time, there will come a time when users will no longer be able to use the old data and the new system will take effect. The change was made in the event "Big Bang". The Big Bang migrations typically have significant setup times and short downtime for which systems are not available [4]. The ideal Big Bang migration has no downtime, but this is not the case for every project.

Trickle Migration takes a step-by-step approach to data migration. Instead of completing the entire event in a short amount of time, Trickle migration operates the old and new systems in parallel, and data is migrated step by step. This method essentially provides zero downtime that critical 24/7 applications need [8]. Migration can be implemented with a real-time process to move data, and this process can also be used to maintain data and bypass future changes in the target system. Adopting the Jet approach adds complexity to the design as it should be possible to track what data has been migrated. In the context of system migration, this can also mean that the source and target systems run in parallel and users have to switch between the two, depending on where they need the information they need. Alternatively, the old system may continue to function until all migrations are complete before the user switches to the new system. In this case, changes to the data in the source system should cause the associated data records to be migrated again so that the target is updated correctly [5].

3 Methodology

The methodology used is the commonly known waterfall model (Fig. 1) where it is broken down into 5 phases. Not only the model is easy to understand but it also provides a better way to present a documentation of a project development. For a data migration project, this is the best model to reflect the details and explanation of each process that have been taken throughout the project.

3.1 Planning

In the first phase, we must define what is possible in theory and reasonable approach to solve the problem. From the organization perspective they cannot afford to spend high budget on data because it is not their core business. Therefore, we turn into open-source software to help with the data migration. Not only it will speed up the process, but also help us to monitor the database. But it is not only the case, open-source software usually has locked content for free users where the tools provided is limited, thus we must come up with other way to fill in the weaknesses. Data migration projects needs to be executed carefully by prioritizing the required data to run the system effectively. Considering the best software to use is also one of the factors for a successful data migration.

3.2 Analysis

The purpose of the analysis phase of a data migration project is to identify the source data to be extracted from the source system and then modify it so that the subject or data type fits the new system and is loaded into the new system. In a data migration project, analysts

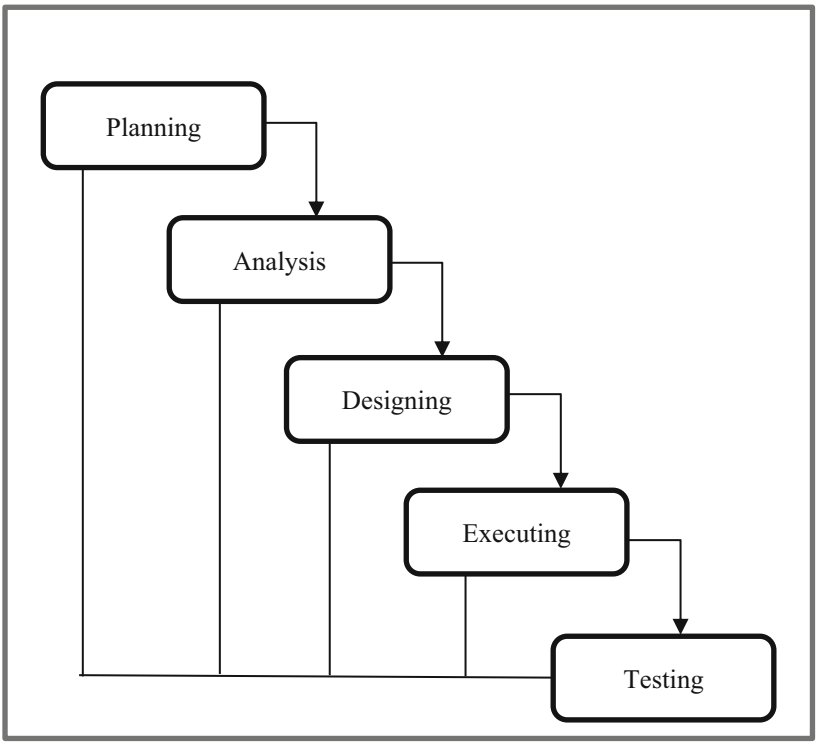

Fig. 1. Waterfall model

identify the mappings between source and target data models at a conceptual level using informal textual descriptions for the ETL process where it must be correctly stated on which data that needs to undergo the process [9]. At this stage we need to consider whether the data quality is good or not. The most common solution is to migrate the source data to a new platform whose data structure is identical to that of the source system. This allowed us to shut down old systems and issue new ones with confidence and without losing historical data [6]. Moreover, it provides clear visibility and access to all data issues with the ability to investigate anomalies at any required depth. In addition, inconsistencies regarding the scope of the data to be migrated are eliminated and their impact on the overall migration project assessed. Therefore, the best strategies to use is the trickle migration where it meets all the criteria for the project.

3.3 Designing

The flow process is shown in Fig. 2. The process starts with extracting data from old AKK System's Database and map the schema with the new database schema before executing the data migration. Unsuccessful migrations can lead to inaccurate data containing redundancy and unknowns. This can happen even if the output is fully usable and adequate. In addition, any existing problems in the data source can be amplified

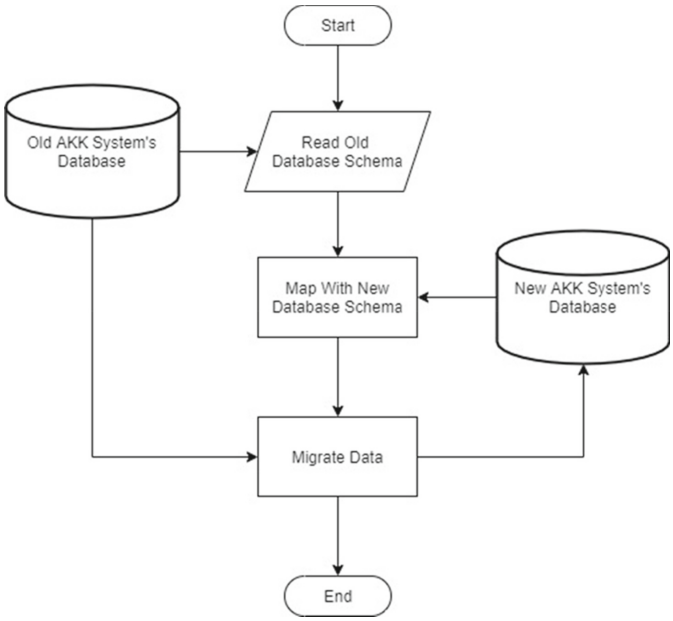

Fig. 2. Flowchart for data migration

when incorporated into new and more complex systems. Common data migration strategies prevent low-level experiences that end up creating more problems than they solve. Incomplete design can lead to a complete failure of the migration project (Fig. 3).

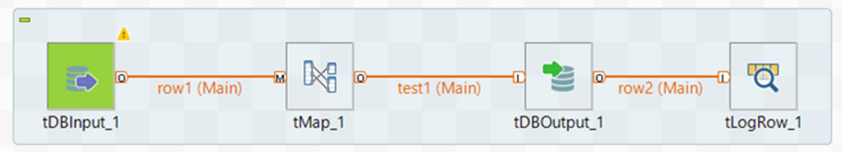

Fig. 3. Talend open studio models for failure

In the next step, is to create models using Talend Open studio (TOS) workspace. Chosen components will determine how the product looks after running the job. 'tDBInput_1' reads the database and retrieves fields based on the query [7]. It executes queries against the database in a strictly defined order which must correspond to the schema definition. The "row1(Main)" link then takes you to the list of fields for the next component. Next, 'tMap_1' is an advanced component that is integrated in addition to the TOS. It converts and routes data from one or more sources to one or more destinations where in this case it is linked via 'test1(Main)'. 'tDBOutput_1' writes, updates, makes changes or suppress database entries. It takes the specified action for the data in the table, based on the flow incoming from the preceding component in the Job where it linked

via ‘row2(Main)’. Finally, ‘tLogRow_1’ Displays data or results in the run console and is used to monitor data processed.

3.4 Executing

Once the model has been created, data is extracted from the source system, modified, prepared, and loaded to the target system according to the migration rules. At basic run tab, there is a route run button and clicking it will execute the job according to prior setting. the console show progress in implementation. The log contains each error message as well as the start and end messages. It also shows the output for the operation if the tLogRow component is used in the job design. Talend Open Studio offers a variety of information functions that are displayed while running the framework, such as statistics and tracks, which make monitoring and debugging easier (Fig. 4).

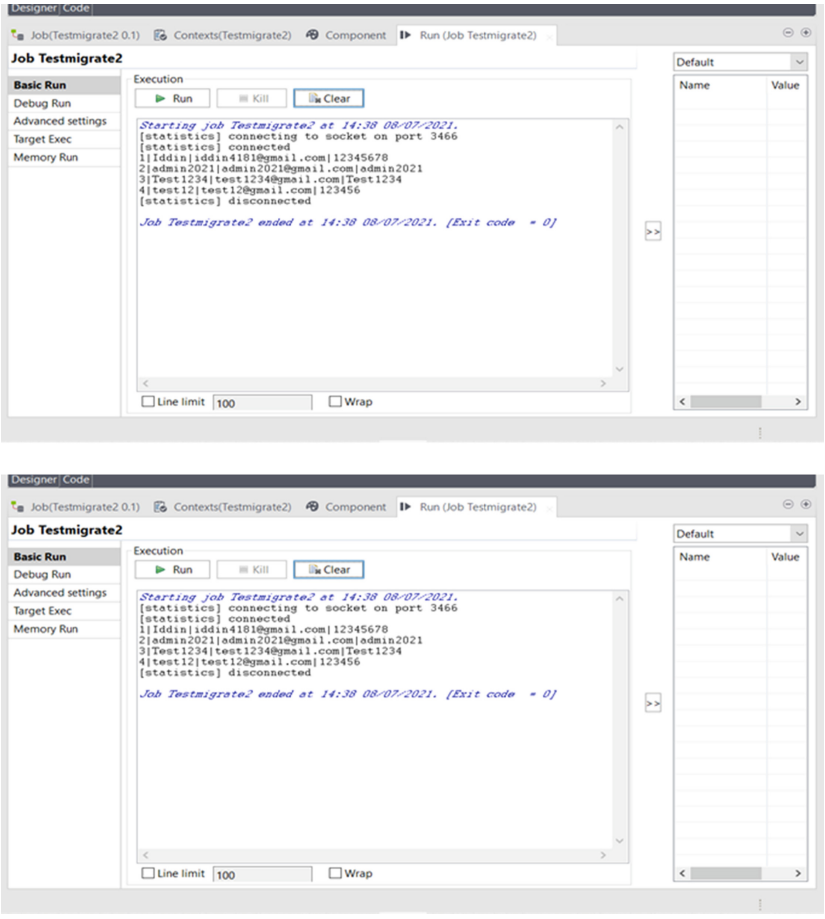

Fig. 4. Job execution

3.5 Testing

Finally, the testing phase requires the model to be tested to see the quality and integrity of the migrated data. The test will be performed on a platform that has similar setting to the actual system's database to increase the rate of success, especially when implement it on a live system (Fig. 5).

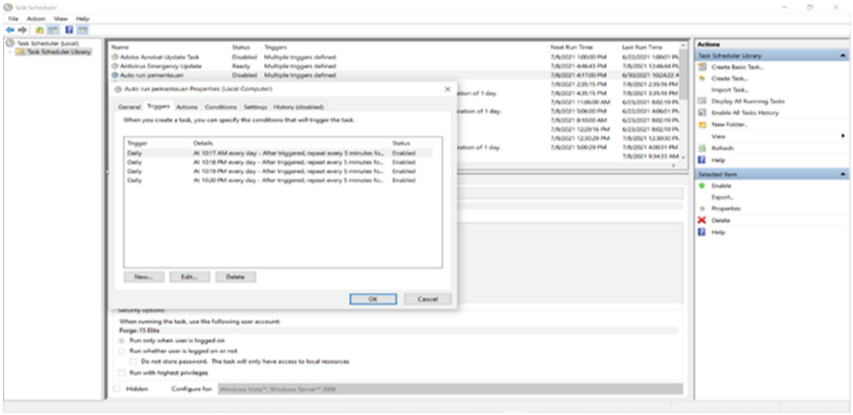

Fig. 5. Windows task scheduler

If the test is successful with no errors or anomalies, the next step is to create a standalone job where it can automate and execute the job on a schedule. This can be done with the help of Windows Task Scheduler. The actual system needs to be updated every 1 min but the task scheduler can only run the job at every 5-min time intervals. This can be solved by navigating to triggers tab and create 4 triggers where each one of it is set 1 min late than their predecessor. After 1 min the first trigger ran the job, the second trigger will run the job again and the cycle continues where it will loop back at the first trigger after 5 min. Therefore, the scheduler can by past the 5-min minimum limit.

4 Result and Discussion

According to the phases stated in the methodology section, every finding is recorded and explained in this section.

4.1 Effectiveness of Using Trickle Migration

Throughout the experiment it is not as smooth as expected. Even if the theories behind the strategies match the main problem, there are other factors that also contribute to the success of the data migration. In this case, the data from the older system during their service is not well managed by the organization because of there are no employee who hold the position to overlook the qualities of the data gathered. Thus, creating more

complication for the project. Therefore, we had to separate the usable data to avoid 'blanks and redundancies before executing the data migration. At the end, we had to do data preparation and validation first where the data is transmitted by batch to a newly database before it could be extracted into the new system to eliminate the uncertainty. From the experiment, we have observed the effectiveness of the Trickle Migration. It is as follows:

1. Zero downtime required.
2. Start old and new systems in parallel and transfer data in small steps.
3. If the single-phase phase fails, only this failed phase needs to be rolled back and repeated the process again.

4.2 Talend Open Studio and Automation Performance

After finishing the testing phase, we find that TOS software is capable to solve most of the problem besides capable to adept with any data migration strategies. It offers not only tools such as cloud, big data, enterprise application integration, data quality, and master data management, but also a unified repository for storing and reusing Metadata. In addition, the Talend ETL tool improves the design efficiency of data migration jobs by setting and configuring them in a graphical interface where we can simply drag and drop the necessary component.

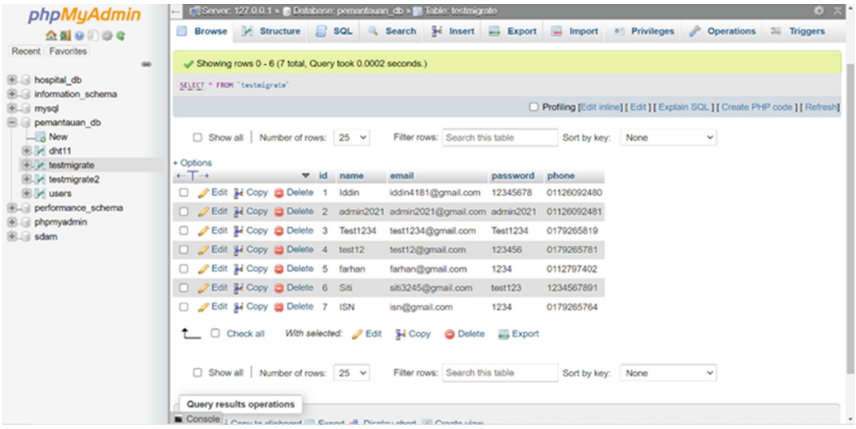

Fig. 6. Talend open studio output

Figure 6 is the output when the job is executed for the first time on a similarly structured of the new system database. This is considered a success when it can extract data from other database then transform the data to match the destination database and lastly load it to the destination database. When new data are inserted or updated, the destination database also updates at the same time. Next, there are no noticeable anomalies, blank space, or redundancy after the job is complete.

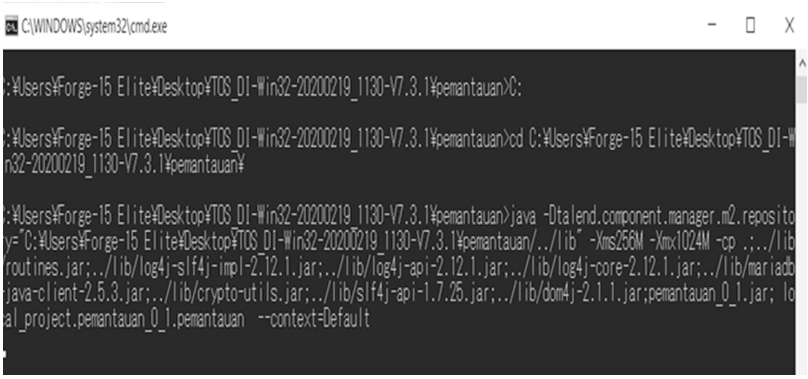

Fig. 7. Window command prompt automatically execute data migration

Figure 7 is the windows command that pop up every 1 min after the job is executed. The finished design in TOS is converted into.bash file format where it can be executed like a normal application. To automate the execution of the file, window task schedular was used and the result is similar to our main objective.

5 Conclusion

In conclusion, data migration can improve an organization performance and deliver advantages in their industry. In terms of efficiency, it is far better than transferring the data manually when technology can do the labor of data migration automatically. Generally, to move all the foundation data structures in the source system to the new system is time consuming where the plan needs to be laid out carefully to get the best result. At some point in the future, there is a high chance for organizations that use IT system will need to transfer a huge set of data from one platform to another. Therefore, it is a mistake if data migration simply considered as part of a larger project. Some organization might be convinced to handle it by themselves, but this can lead to decreased efficiency and increased costs down the road. If an organization is planning to migrate large sets of data, it's essential to do research on the best practices for a successful data migration.

Even with the right data migration tools, the migration process can be long and arduous. By educating oneself on the why and how of the process, it will help deciding what type of data migration the organization needs to undergo. Having a reliable team of experienced employees and consultants can help an organization ensure this process goes smoothly.

References

1. Lelii, S., Hefner, K.: What is data migration? - definition from whatis.com. SearchStorage 28 April 2017. https://searchstorage.techtarget.com/definition/data-migration
2. Data migration: Strategy and best practices - talend. Talend Real-Time Open-Source Data Integration Software, 2 October 2020. https://www.talend.com/resources/understanding-data-migration-strategies-best-practices/

3. NetApp. What is data migration? – how to plan a data migration. NetApp, 1 January 1 1970. https://www.netapp.com/knowledge-center/what-is-data-migration/
4. CloverDX. (n.d.). Data migration. CloverDX. https://www.cloverdx.com/explore/data-migration#:~:text=A%20'big%20bang'%20data%20migration,single%20'big%20bang'%20event.
5. Oracle Successful Data Migration, October 2011. https://www.oracle.com/technetwork/middleware/oedq/successful-data-migration-wp-1555708.pdf
6. Joseph, R.H.: An Overview of Data Migration Methodology. An overview of data migration methodology, April 1998. https://dulcian.com/articles/overview_data_migration_methodology.htm
7. Talend Open Studio Components Reference Guide. Welcome to Talend help center. (n.d.). https://help.talend.com/reader/hCrOzogIwKfuR3mPf~LydA/_KmjKFtYUBb9Wd_P~whPcg
8. Nyeint, K.A., Soe, K.M.: Database migration based on trickle migrations approach. Natl J. Parallel Soft Comput. 81–86, (2019)
9. Yeddula, R.R., Das, P., Reddy, S.: A model-driven approach to enterprise data migration. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 230–243. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19069-3_15