International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

RPA to Automate ETL Processes in Data Warehousing

Tapan Kumar Rath[1], Sibaram Prasad Panda[2]

Email: *tapankumarrath001[at]gmail.com*

Email: *spsiba07[at]gmail.com*

Abstract: ***Robotic Process Automation (RPA) is an automation technique that utilizes scripts and robots to execute repetitive tasks in a business process. RPA can free up humans to perform more logical tasks that require reasoning, judgment, and emotions, while robots are programmed to do mundane and monotonous recording and processing tasks. RPA can perform these processes with a high level of accuracy and less time. RPA is implemented for automating the repetitive tasks performed in enterprise applications. RPA can be applied in areas such as finance, management, cybersecurity, healthcare, travel, and logistics. RPA finds its applications in sectors such as record keeping, compliance, cyber security, technical support, and more services. RPA provides faster and more reliable processing of the Automation at a reduced cost. Most of the developed companies have started investing in RPA. There is also a huge market potential in RPA.***

Keywords: Robotic Process Automation (RPA), Extract-transform-load (ETL), OLAP cubes, OLTP schemas, cyber security, Mapping Data Flows, CRM

1. Introduction to RPA and ETL

Data for Business Intelligence is derived from unique sources whose data needs to be automatically pulled out through RPA instead of manual intervention. In ETL, the data from various disparate sources is extracted, cleansed, if necessary, transformed into the required format or structure, and loaded into a Data Warehouse. An ETL/ELT process must and does happen in-process. Extract-transform-load (ETL) is a special case of data handling prior to bulk load of a target database. ETL is performed for data manipulation exploiting all of the benefits of consistent connectivity and transactions possible in process, as well as any source-side filtering. RPA is a technology or software designed to reduce the number of hours or costs an organization runs in performing repetitive or mundane tasks in business processes.

2. Understanding Data Warehousing

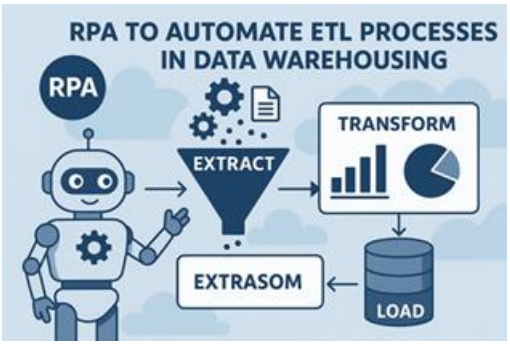
Data warehousing is mistakenly perceived by some as a generalized term for data storage management and is most often combined with other words forming phrases like "data warehousing architecture", "data warehousing design", or "data warehousing online analytical processing applications". However, the term data warehousing in its strict sense implies a special environment devoted to data analysis and different from the operational one, with a unique relational database architecture that is independent of the operational environment with no query-processing overhead, and where an unasked demand can be afforded at low cost. A data warehouse is a repository whose size is measured in gigabytes and whose content is in a structured format available for fast access. Denormalized star schema or Snowflake schema relational structures powered by a powerful database management system on top of a fault-tolerant data platform meet the requirements imposed to a data warehouse and its use.

The data warehouse is the foundation for enterprise data. It allows company-wide information in a design that is more predictable than those used by operational databases. Data in a warehouse is available to those who need it, even if the departments who use the information are not involved in its original creation. From the data warehouse, users can readily analyze data from different points of view. Furthermore, the data are at a lower cost than that required by the operational systems because of the lack of space overhead for the large number of indexes, the validation of the data through their original creation by the systems that update them, and the optimization for a unique use of query-intensive data by the warehouse specialized design.

3. The Role of ETL in Data Warehousing

From a high-level perspective, businesses process data to achieve certain objectives. Data governance is usually in place to afford and ensure that objectives are met. Therefore, corporate databases grow organically over time with different businesses capturing data that realize some micro objectives. As time progresses, the data within these databases become stale and demand for new insights together with advances in technology requires that businesses take a cross-sectional view of widespread goals. This requires that they have access to high-quality, readily available data. Data warehousing provides businesses access to high-quality, readily available data.

Extraction, transcoding, and loading are phases of the data warehousing process. ETL tells us a lot about the data: what type of data is in the warehouse; the integrity of the data; its relation to real-world constraints; the decisions that went into choosing which source-level files to include; the logic behind the SCRs; and importantly, how the design will facilitate or inhibit loading and maintenance of the warehouse. ETL is the foundation of the warehouse. How good the data is in the database determines in large part the level of richness of analytical insight and intelligence derived from the data warehouse.



4. Overview of Robotic Process Automation (RPA)

A widely known form of process automation is Robotic Process Automation (RPA), which offers benefits both for businesses and customers. In simple words, the RPA can be stated as the use of automated tools and software, sometimes called bots, to mimic human actions and assist interaction with digital systems using structured data. Structured data are information that resides in fixed fields within a record or file. The potential to deploy robots doing repetitive work around the clock, every day of the week, has generated substantial interest in RPA. There is a huge economy of scale by automating lower-skills jobs. Most hard business cases for RPA focus on middle-skill jobs. Outsourcing previously low- or middle-skill processes, then automating them, often flatters corporate financial performance at the expense of worker security and community stability. The equally huge economy of labor jumps when AI lessons, especially machine learning, are added to RPA. RPA works by essentially elbow tapping employees across the organization to take on the manual workloads best suited to automation. Deploying these bots throughout an organization allows businesses to reduce headcount, speed up service offerings and upgrade customer experiences without dramatically increasing costs. RPA is the gateway to AI.

RPA uses technologies and software that assist in connecting to and transmitting information between multiple disparate systems on a user interface level. This bot creates a bridge between existing systems and mimics the actions that humans perform on those systems to create data transactions across systems. Since RPA is deploying software that emulates tasks that humans do currently, it does not require the investment of time and money required when creating applications and APIs. Instead, users can begin development and deployment in days or weeks. The company that implemented the RPA experience story described at the beginning of this essay used external contractors on a time and materials basis and was able to build applications in only nine months.

5. Benefits of Automating ETL Processes

Data extraction, transformation, and loading processes are often considered the most critical and time-consuming activities in implementing any data warehousing project. Most of the ETL operations are unpredictable and require significant manual effort for implementation. Other risks associated with traditional manual ETL operations involve reliance on specialized engineers and complex technologies to create and manage repeatable data processes. Enterprise application data sources are increasing. This results in more complex data loading which includes hundreds or thousands of different data mappings that must be defined, managed, and transported from source to target. Looking at these challenges, typically organizations adopt ETL automation techniques.

Data extraction, transformation, and load automation has been a growing focus area in enterprise data management since the early days of data warehousing, but technology has matured, and the benefits of ETL automation to the business have become more concrete. Automating the data extract, transform, and load process alleviates the reliance on cookie-cutter, manual coding of ETL tasks. With the complexity of some enterprise data loading operations, automating data mapping simplifies the process and reduces costs significantly by shortening the time to develop and deploy the solution. ETL automation can reduce the need to create a costly custom script for many data mappings and help organizations get more from their technology investments. The result is that ETL automation can bring organizations lower costs for the same data management tasks. This includes lower development and deployment costs as well as reduced maintenance, support, and operations costs. Companies typically see a reduction in resource utilization for efforts involved in enterprise data loading. These savings translate into faster availability of information for business users.

Automating ETL processes can deliver business advantages such as better data, better information, and faster processes. Automated data quality checks can eliminate unnecessary rework resulting from bad data. Higher levels of trust and ownership by the business lead to better performance. Faster, near real-time data availability can open up new business opportunities. But it is important to note that although automation helps, it isn't a magic cure for the pain of data integration.

6. Challenges in ETL Automation

Complexities arise during the design phase of an ETL process because a deeper understanding of all the data sources and destinations, as well as the trustable data mappings, is required to integrate the data sources. ETL operation requires different reasons for the different steps to be defined and documented. Security, compatibility, and multitudes of data sources must be considered. Automating ETL operations projects can be complex, especially in organizations with large, heterogeneous infrastructures. Complex issues regarding security, access, recovery, and file formats must all be addressed before implementing a successful automation infrastructure. The tools require experience, skill, and an understanding of the ETL requirements to implement the automated infrastructure successfully. Implementing an ETL automation infrastructure is an art unto itself. Below we summarize the core issues and steps. There are many possible issues and questions relating to the automation of the ETL process. Each organization must address its own set of issues. ETL automation should not be a box of tools used willy-nilly by anyone responsible for ETL.

7. Key RPA Tools for ETL Automation

RPA tools for ETL automation allow organizations to design new ETL processes or enhance existing ones without costly custom code. RPA tools are popular for ETL automation. One tool has a collaborative Data Gateway for seamless collaboration between business users and IT. This data integration service allows IT to build a Data Gateway that orchestrates communication between systems and various file types. It enables users to leverage the tool to deliver the data integration requirements for business processes across any system for sound decision making. Bot-Agents execute the data flow orchestrated in the Data Gateway, while IT defines the data source and destination connection criteria.

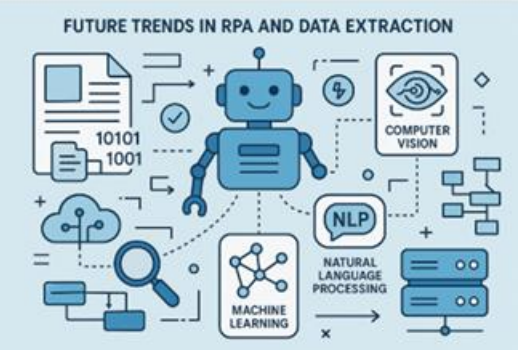
8. Designing an RPA Solution for ETL

An RPA solution for ETL is intended to minimize overt automation, with the majority of employable ETL tooling being vendor driven. When looking for workarounds to ETL-toolset limitations, it's best to avoid waiving qualifications for RPA-like behavior with a single-machine, because RPA works across machines with little IT coordination. Similarly, support from multiple vendor tool-chain packages ought not to make RPA work harder than it has to. The RPA implementation writer does no ETL magic with bots, but sticking to thinking like an RPA developer defines what should and shouldn't go into the scenario. Even with limited support from the ETL toolkit, there are advantages to looking ahead to.

8.1 Identifying ETL Processes for Automation

RPA enables new approaches for undertaking ETL processes in Data Warehouse environments. The new approaches are enabled by three capabilities of RPA. The first is RPA technology's exceptional competence in dealing with unimodal, unintegrated systems. Second is RPA technology's unique capability to undertake non-intrusive, point-to-point integration between different and disparate systems. Third is RPA technology's integration of business process management features within it. These three capabilities of RPA can be leveraged to address the two key disadvantages of existing ETL technologies. The two disadvantages are the relative complexity of setting up rules and interfaces in existing ETL technologies and the generalized, pre-programmed nature of the existing ETL technologies. Since RPA is business process driven and easy to customize for ad-hoc processes, it can complement existing ETL technologies to fill gaps.

The first task in the design of an RPA solution for ETL is to identify ad-hoc processes, short-term batch processes and point-to-point integration needs. Identifying needs for ad-hoc processes is usually not very difficult. However, identifying needs for short-term batch processes and point-to-point integrations is more difficult. RPA tools can



fill gaps in both ETL and ELT modes. Companies usually undertake traditional ETL using an ETL engine. ETL is made up of extract and load from original data sources that reside in non-database applications and that also may reside in databases that do not have data adapters built for them in the ETL engine. In the ETL engine, extract also may deal with data that requires business rules to convert the data into valid format to meet data quality needs. These databases need to relate to the ETL engine through custom code. After undergoing transformation in the ETL engine, the cleansed data is then put into the data warehouse either in an operational database or an analytics database. The analytics format is usually columnar and is optimized for analytics queries.

8.2 Mapping Data Flows

An end-to-end ETL design is required to specify the usability and functional requirements as well as the packaging required to safely and effectively deploy the solution into a production environment. For other types of software systems and solutions, the design often involves high-level block diagrams, then develops more detailed diagrams, followed by pseudo-code before starting construction. RPA tools are designed to leverage business and process knowledge, engaging process owners during all phases of the solution lifecycle. For RPA, however, the majority of these steps are performed at the same time and at the level of detail required for efficient construction, testing, and operation.

9. Implementation of RPA in ETL

As ETL automation using RPA is a new area in Data Warehousing, there is very limited literature on this topic. The use of RPA technology to speed up data preparation by automating repetitive data tasks should be investigated in more detail. A step-by-step implementation guide based on these general steps of the RPA development life cycle includes: (1) structure the processes, (2) design the process, (3) build the process, (4) test the process, (5) deploy the process, (6) manage the process, and (7) improve the process. These steps are followed to implement RPA in ETL in this order, with general definitions listed as follows.

- Structure the Processes: Identify all simple and repetitive data tasks required to carry out ETL in a low volume of data at a low and manageable level of accuracy to establish a list of possible tasks to be automated. List these tasks in a table with attributes such as frequency, duration, required frequency for review, bottleneck operation in a simple ETL process workflow, estimated time for the Robotic Process to run, and easiness rating. By using the attributes discussed, take the top data tasks in the output table as candidates for implementation.
- Design the Process: Identify the specific tools to be used for RPA and design the robots with associated workflows to carry out ETL. Specify the information that bots need and how they will share that information, along with the milestones for task completion.
- Build the Process: Use the tools identified in the previous step to develop the Automation Robots and run the test routines of the Robot.

9.1 Step-by-Step Implementation Guide

This chapter describes how to do RPA in ETL in three major steps: Assembling RPA Infrastructure, Designing Process Maps and Creating RPA Bots. Those are the three critical activities of any RPA in ETL implementation. We elaborate on these major steps in the sections below.

1) Assembling RPA Infrastructure. This step involves selecting the right RPA platform for ETL automation and building RPA capability using partners, people and pre-built components.
2) Selecting RPA Platform for ETL Automation: The first step in this process is to identify the specific vendor tools that can service data warehouse ETL activities and analyze their specific capabilities. You will need the RPA vendors' support for this initial process, given their tool breadth and strengths in different use case areas. This selection process should ideally involve the team designing the RPA in ETL solution. Additionally, tasks that could not be automated using traditional RPA should be able to support automation using APIs that would meet the security and data use case objectives of the enterprise.

9.2 Testing and Validation

Before the deployment of any RPA process, it is recommended to perform both User Acceptance Testing (UAT) and validation. UAT is primarily a verification process to check whether the new RPA process meets the business needs and requirements for which it is being automated. It validates the outcome of key process activities and ensures that process input created by a business user is giving the expected output. There are no fixed rules to define the duration of the UAT but, generally, it is advisable to carry out UAT for the anticipated number of variations in the process to be automated with a good mix of average, worst, and best-case scenarios. The activities executed during a normal business day by the customer domain are good candidates for performing User Acceptance, it is suggested not to touch upon either the payroll or the data until the process is completely stable.

10. Case Studies of RPA in ETL Processes

In this section, we describe two case studies of commercial process automation addressing the ETL processes in the industry. The first example is a bank that developed ETL processes in Data Warehouse operational management. The second example has two ETL processes of one retail chain that provides integration and aid in business Intelligence and Control.

10.1. Industry Case Study 1

This project in a large bank developed three process automations that generate ETL Data Warehouse management reports: AB_MA_DPR, AB_MA_FIN and AB_MA_PR. The Data Warehouse supports the operational management of its various departments: Asset Management, Economic & Financial Policy Information, Financial Institutions and Internal Treasury. Reports are executed daily, weekly, and monthly, and bring together information drawn from various data sources that consolidate and overlap Data Warehouse bank databases.

Our Approach: Review ETL processes that generate the bank Data Warehouse management reports. Process the ETL process flows through an RPA tool. Specify the RPA information processing robot. Deploy the process robots in Data Warehouse scenario. Validate the deployed robots with operational management users.

Results: Considering only the AB_MA_DPR Process, we have offset the processing time by more than 98%. Before RPA, the execution time of that Data Warehouse Management report was more than 64 hours. After RPA, we processed it in less than 1 hour, and we just changed the report execution frequency from weekly to daily. Processing was performed in a special Data Warehouse area, created for that purpose, in Off-peak hours.

10.2 Industry Case Study 2

This project in a retail chain has two RPA process automations that provide integration and assist in Business Intelligence and Control, operating in the Commercial and

Price areas. Such processes are potentializing candidates for RPA mainly because they work mainly with spreadsheets, which normally have a very large volume. They are executed weekly, have a long execution time, filling various reports and spreadsheets that would take a long time to fill manually. Also, the analysis is unnecessarily delayed.

10.3 Industry Case Study 1

In this section, we present use cases of RPA to automate ETL processes in data warehousing. The first is a business case of a multinational corporation firm, while the second is a business case of an open-source cloud software service firm. We used these business cases as an inspiration to conduct our study. Located in France, the multinational corporation firm is specialized in urban and transport sustainable mobility solutions, telecommunications, energy, safety, security, etc. In this case, we demonstrate how to semi-automate a BI ETL task to distribute a report to a predefined list of submitters on an individualized basis. The Data Visualization team at the corporation has developed “weeklies" of data to serve as operational dashboards for various stakeholders of the Group, both at the headquarters and in the subsidiaries. These “weeklies" provide a view of research and development costs incurred, billed, and related to the number of people assigned to each project. The “weeklies" are distributed to senior managers on a weekly basis. RPA is used to semi-automate the ETL task. The software automates the process of completing and distributing the “weeklies." The Robot is triggered every Thursday night and connects to an analytical native Salesforce BI to retrieve the data for the report. The data is then stored in an SQL database. The report is produced in PowerPoint based on a template that follows the Graphic Standards of the Group. The report is emailed to the predefined list of submitters. Finally, a log file is created to ensure that the Robot works properly throughout the operation.

10.4 Industry Case Study 2

u0170. - 10.2. Industry Case Study 2 This case study describes the use of RPA to improve ETL processes for a leading telecommunications company called "Telecom". Telecom´s people faced challenges with data ingestion from diverse and large - volume sources, including files in SAS and CSV formats, data ingested by third-party systems via web services, and data stored in hardcoded Oracle tables. The data assimilation consumed a huge effort spent on the repetitive and manual steps, delaying the availability of timely and trustworthy information. Failure to timely deliver trustworthy reports required constant and expensive human intervention in related business processes. Needs for arduous but redundant work required recruiting extra business intelligence professionals, increasing the total cost of processes that were already deemed critical for the business. Telecom started exploring ways to better manage their ETL operations when the quality of service became unsatisfactory. Telecom decided to implement RPA using its Operating Center to run a set of automated tasks developed by an external company specializing in process automation. The tested pilots involved a set of ETL scripts created based on Telecom´s ETL prior work, executed manually or using scheduling tools by Telecom´s professionals, and they mocked the output of the automated tasks when they implemented the process manually. The company automate of automated task, however, was capable of finishing its service in shorter periods of time. Time savings helped Telecom to deliver more timely and more reliable information to business managers on a variety of business-critical analytics, including forecasting of bank service demand and revenue from demand by bank service. The automation project of the ETL processes became applied to other Telecom internal and external processes and generated operational and customer experience benefits to the business.

11. Performance Metrics for RPA in ETL

It is crucial to have a metric framework for RPA to automatically track and measure the value, productivity, and efficiency of every RPA automation that runs in near real time. With significant funding and high expectations for speedy RPA deployment in organizations, it is increasingly pertinent to ensure RPA is delivering against its targets. Key metrics can help to manage RPA expectations, avoid the fatigue and disappointment of failed RPA projects and ensure RPA delivers more value for the resources and investment utilized. A metric framework for RPA provides key stakeholders in finance, risk, governance, and audit with a comprehensive overview of all governance and risk around RPA initiatives. Forward-looking metrics identified in collaboration with business leaders can also provide vital insight into any expected business process changes that can significantly impact process efficiency. Using these insights, resource allocation and support can be organized to ensure RPA is given every chance to perform.

12. Future Trends in RPA and Data Warehousing

With the advent of Artificial Intelligence and Machine Learning, the future of RPA is assured. Although the basic tenets of RPA remain, the addition of intelligence and cognitive capabilities for execution, supervision, and monitoring mean that RPA would operate closer to being a plug-in for execution for Intelligent Automation. While these new intelligent capabilities would have application across multiple functions, the role RPA plays in ETL processes is significant. The execution of ETL often considers a multitude of diverse factors. It gets data from multiple systems. It requires flexibility and ease of configuration. It is but a cog in a bigger wheel. Business functions driven by ETL processes often work at various stages of maturity

13. Best Practices for RPA in ETL Automation

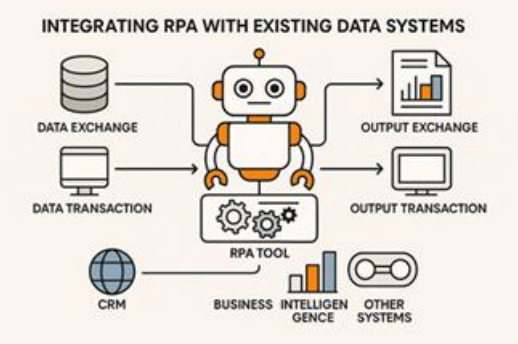
RPA makes it possible to automate ETL processes without having to invest in more complex ETL tools. However, RPA for ETL is designed to work in a particular way. If the design, implementation, and management of ETL automation with RPA is not done in a certain way, it is possible that it will not work effectively, or quickly enough, to be worth the effort. There will, of course, be differences in how content is implemented depending on the RPA tool you are using, the bots or processes you have created and how you are using them, and the ETL process you are automating. The simplest

ETL processes are the best candidates for initial ETL RPA implementation. ETL processes that are extracting data from and loading it into standard flat file formats such as CSV or tab-delimited files are easier for RPA tools to accomplish. ETL processes that need validated data arriving at the data warehouse tables by a specific time, such as every hour, day, or week, are the best candidates for RPA implementation. With further implementation or custom-built bots, those processes that take hours each night to run may also be suitable for RPA implementation. If the ETL process you want to automate requires an agent and/or uses complex database queries that need to be created, edited, or modified regularly, RPA implementation will likely add more work than it saves. Most RPA tools have a separate interface for validation of input and output files. Building that interface requires additional steps not required with other ETL tools. And RPA tools may only allow certain file formats as input or output.

14. Integrating RPA with Existing Data Systems

How does an RPA tool make decisions at any step of a bot run? It is not maybe possible to capture every nook and corner of the IT landscape but RPA by nature is exploratory and customized. The organization needs to partner with vendors that have capabilities listed above. They have predefined connectors on their toolkit or extensive industry experience. They allow bots to connect with the input environment at the data exchange element or data transaction element and the run-off environment at the output exchange element or output transaction element. Organizations can also develop custom libraries for individual systems, which means a temporary maintenance headache.

Once the bot is connected to the left and right systems, the RPA tool can mimic end-user actions to accomplish tasks such as logging in, navigating screens, entering data, performing calculations, running reports, and generating output, whether populating data back into a system or generating reports. Middleware vendors hence are rapidly becoming important partners for larger digital business automation players. These partnerships expand the number of applications that can be integrated with RPA. As a result, what this means practically is that organizations can collect business data from one or more sources, then carry out data operations and then push the outcome into one or more back-end systems or tools.



15. Security Considerations in RPA and ETL

Security is at the core of all RPA implementations, and these include projects for automation of ETL processes that transfer organizational data and information between systems governed by the ETL processes. Data privacy, security, and governance become crucial to ensure organizations do not compromise sensitive company data. RPA implementations must follow security controls from organizational security policies as well as the guidelines for data security. Organizations must take extreme care to protect the credential management required for RPA scripts since all sensitive data may be exposed via credential management.

Whenever possible, data used during automation with RPA should be encrypted to ensure sensitive data is not visible in the RPA log files. Organizations must account for the cost of secure credential management when choosing to do RPA implementations. In most cases, credentials must be stored in a file or memory, which RPA vendors provide methods for these storage mechanisms and adequate logging. All security regulations surrounding ETL activities should be observed for RPA implementations for ETL. Data latency, data sub setting, data masking, and data encryption should be used in RPA implementations when it is not necessary to perform a particular ETL process on masked data.

Testing of all security-related functionality, such as authorization, traceability, data encryption, and error handling should be included in the RPA scripts testing phase. The security of business functions scripted to be performed remotely via RPA implementation should support all security functions required by the enterprise application, such as multifactor authentication, certificate authentication. Security configurations for both the enterprise application as well as the RPA script must adhere to security guidelines for secure operations.

16. Regulatory Compliance and RPA

A necessary condition for properly meeting the mission of intermediaries is compliance with applicable regulations as promulgated by regulatory authorities. The nature of regulatory compliance varies substantially for the different types of intermediaries. Since such differences relate to the actual activities developed and the actors involved, and RPA is inherently flexible, appropriate implementation of RPA in diverse types of intermediaries can significantly improve regulatory compliance and consequently ultimately mission integrity. In the case of banks, implementing RPA enhances compliance with anti-money laundering regulations. The implementation of RPA in accounting service firms significantly moderates the risk of breach of the regulatory compliance framework when conducting audits of public entities, enhances their commitment to integrity assurance of the audit process, and consequently mitigates their liability as gatekeepers in the financial system.

Enhancement of their activities as intermediaries, in turn serving to facilitate regulatory oversight and enhancing the integrity of the overall system of which they are a part.

17. Training and Change Management for RPA Adoption

The introduction of robotic process automation into an organization requires a shift in thinking for many users, and to facilitate this shift, careful attention must be given to training and change management. Training is an opportunity to help employees understand and appreciate the technology and architecture, what processes are suited to automation, how to identify processes worthy of consideration, the RPA development lifecycle and effort classification, and the limits of the automation initiative. It is used to encourage prototyping and to contribute to the development of automation skills among the early adopter community. Training can help allay fears about automation being a threat to human jobs. RPA adoption represents an evolution rather than a revolution; human labor is still needed for the more nuanced and complex business activities, and those tasks that are being automated are ones that are disliked or looked upon as menial. With the process-building tools being so simple, users of the software can focus their time and energy on the value-adding parts of the end-to-end process of which their operations are just a small part. They can revisit each process and look to improve it. In other words, automation takes away the pain of doing the same repetitive tasks over and over, freeing users' time to work on higher-value enterprise improvements.

18. Cost-Benefit Analysis of RPA in ETL

RPA for ETL has also significant advantages compared to traditional ETL tools, apart from some pitfalls. RPA is relatively less expensive, can be up and running in less time, has a shorter learning curve, is more flexible and adaptable, is more automated and can save time on repetitive human actions, and has lower hidden costs. Because of the advantages of RPA, a different demand has emerged in the ETL automation industry that traditional vendors have a more inferior product for. Because of these advantages of RPA, RPA tools have grown significantly. Popular ETL vendor tools have marketed their RPA products to either assimilate or compete with traditional RPA vendors. Many small niche vendors also provide functionality in that market and would be viable for many ETL automation cases. Automating ETL pipeline using RPA is less expensive, up and running in less time compared to legacy solutions and has a shorter learning curve.

19. User Experience and RPA in ETL

The user experience with traditional ETL technologies can be daunting and frustrating, even with the best set of user functions. Often, it is the unhappy users of ETL technology who decide the ETL architecture. ETL users will take advantage of automation or software products to relieve their own workloads, no matter what the officially sanctioned ETL technology in-house is. Many organizations have established their own teams of users who write ad-hoc utility programs to meet their personal ETL needs, preferring to bypass the official in-house ETL systems. Hence, the mass user acceptance of corporate ETL technology is not just a set of well-documented functions. While a set of well-documented ETL functions will attract and keep the most skilled developers and ETL Architects who do ETL work in-house, the skill set of the regular ETL user base within a company will largely dictate the functionality and conviviality of the ETL technology itself.

20. Feedback Mechanisms for Continuous Improvement

Regular feedback about daily operations is critical to the success of any organization—whether focused on automation or not. Manual processes need regular monitoring, and RPA processes are no exception. Because responsibility for processes using RPA often shifts from the IT team to a business team, many businesses are unaware of the constant support needed. RPA helpdesk operations are typically much smaller in scale than any other automation solution but still require monitoring.

The monitoring of any RPA process doesn't need an FTE, but depending on the number of processes in place, a part-time role tasked with keeping a watchful eye on the project's daily activity is essential. Whether a company has one automation in place or twenty-five or more in various business functions, daily monitoring of both success and failure rates is critical to efficiency. By excluding a monitoring phase after automation is put in place, optimization would never occur within the organization.

Upon diligent review of activity logs and processes in production, patterns will inevitably emerge over time, and once noticed, the changes and improvements can have drastic effects. With such a low and in some cases missed investment in RPA projects and processes, organizations must track activity and efficiency closely and not overlook what's considered the final step—monitoring the health of the automation. Partnering and considering citizen developers in business units where the automation is housed can provide you with opinions on feedback, proper tracking, visibility, and the possibility for adding on to automation tools.

21. Conclusion

In this work, we explore the application of Robotic Process Automation (RPA) to automate Extract, Transform, Load (ETL) processes to populate data warehouses from internal and external databases and flat files that require extensive, repetitive data retrieval and transformation efforts. We

discuss the main concepts, illustrate how to design RPA workflows to implement ETL, and use an example to discuss the pros and cons of RPA for ETL. ETL performance is key to approximating data warehousing technology to transaction processing technology, standard API links to source data are rarely available, and data retrieved across sources must undergo extensive transformations before they can be loaded. The originality of this work is based on the combination of two ideas: First, ETL processes require the execution of complex series of repetitive steps that can be easily automated; second, RPA tools have many ready-to-use utilities, features, and connectors that simplify the automation of these simple repetitive step functions, for example, OCR, Excel manipulations, and database connectors. Any computer user, business or functional analyst, can create an RPA workflow. If there is a bottleneck in populating a data warehouse, often the first and simplest step to speeding the process up is to automate it using RPA technology. ETL is only the first part of the data warehousing process. This stored data have to be kept updated and in sync with their original sources and be constantly refreshed. RPA workflows can perform both tasks: poll the data sources for updated records and refresh them in the data warehouse.

References

[1] Jamaluddin, M. S., & Azmi, N. F. M. (2016). Extraction transformation load (ETL) solution for data integration: a case study of rubber import and export information.
[2] Machado, G. V., Cunha, Ítalo, Pereira, A. C. M., & Oliveira, L. B. (2019). DOD-ETL: Distributed On-Demand ETL for Near Real-Time Business Intelligence.
[3] Pratap Deb Nath, R., Romero, O., Bach Pedersen, T., & Hose, K. (2020). High-Level ETL for Semantic Data Warehouses - Full Version.
[4] Omondi Odero, J. (2019). Intelli Bot data cleaner: a study of Kenya Revenue Authority's data cleaning exercise