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

Optimizing RPA for Intelligent Data Extraction from Heterogeneous Databases

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

[1]Email: *tapankumarrath001[at]gmail.com*

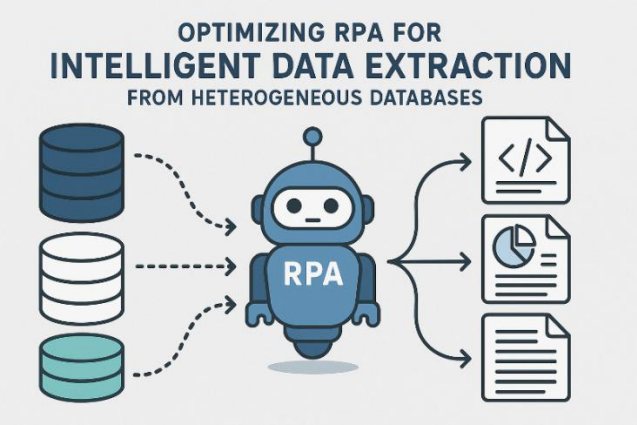
[2]Email: *spsiba07[at]gmail.com*

Abstract: ***We present a method for optimizing Robotic Process Automation (RPA) focused on intelligent data extraction from heterogeneous databases and creating a new record of aggregated information. The databases may differ in structure and in the kind of products registered on them. The method improves RPA efficiency in terms of FTE-hours and runtime. RPA aims to automate routine tasks carried out by humans, thus freeing their time for more creative activities. Robotic Process Automation is software-based technology that allows companies to create automated process flows that interact with different 'Digital Workers'. RPA employs a Digital Worker, called 'the Robot', who is created in the image of the Human Worker, and is programmed to execute the same repetitive route and follow the same rules of Human Workers on Digital Systems. The Employees spend their working hours completing a variety of simple, repetitive operations. The purpose of RPA is to enable the rapid and cheap construction of software tasks, which can be easily deployed and targeted at existing IT Systems, with enough flexibility to be able to take on all different exceptions programmers would normally be needed to cater for through traditional programming techniques.***

Keywords: Robotic Process Automation (RPA), Natural language processing (NLP), Machine Learning (ML), Artificial Intelligence (AI), Transactional data, Relational databases

1. Introduction

In normal days, Human Workers apply the same routinary and boring activities, but during the fiscal close period, they need to put extra effort, dedicate more working hours, and overcome a lot of stress to meet the deadline, and still, they make errors. But what if we create robots in the image of Human Workers of these tasks?



2. Understanding Robotic Process Automation (RPA)

Robotic Process Automation (RPA) is a powerful technology that allows organizations to automate mundane, repetitive tasks, which would otherwise demand resources from the focus areas of the business. RPA can be defined as an enterprise automation solution that deploys software "bots" to carry out high volume structured tasks previously performed by humans, efficiently operating across application silos without changing the underlying system or replacing the existing applications. Since the introduction of RPA as a software-based protocol, most implementations have been internal, largely focused on back-office automation. By mimicking human tasks, the RPA bots read, launch, interpret, and complete these tasks on different applications. There are several advantages that justify the use of RPA to develop automation processes. RPA reduces total costs, increases speed, increases ROI, reduces errors, enables faster onboarding, and frees up internal resources. RPA should be on any enterprise's roadmap to modernize and augment business processes, however over time AI has the potential to replace many RPA tasks. And there are already evident trends that are driving Intelligent Automation (IA) development. Second generation RPA will run alongside Cognitive/AI capabilities to process unstructured data, like emails, social media, and review complex financial transactions requiring decision-making and judgment. Organizations are already exploring a broad range of new use cases for incorporating Intelligent Automation (IA) alongside RPA tools to add capabilities and improve processes. RPA and IA will extend the reach of both AI and RPA technology, processing more varied data types to deliver enhanced outcomes, while bringing in additional IT partners to automate, innovate and truly transform organizations.

3. The Role of Intelligent Data Extraction

Data is a valuable resource for driving the progress of enterprises. Organizations utilize a multitude of structured and unstructured data, which, if adequately processed, transformed, and made available within the enterprise, can lead to business success.

Modern exploitation and Big Data Analytics and such digital business models demand a more flexible and adaptive approach to non-standard data composition and integration. These processes might need to be intelligently triggered by users or user roles disposing of some authority, in the context of specific operational processes, in order to leverage data by elaborating ad-hoc queries and visualization tools, ultimately leading to data-driven actions and decisions. More than just a

dynamic and on-demand policy enforcement, this concept assumes the ability to leverage data in an ad-hoc manner using less-technical users, who are capable of recognizing the data element needed to create or to fill the information gaps they notice in their operational analytics.

In this respect, intelligent data extraction interfaces can significantly help simplify and automate the interaction between users and databases by querying or even mining data residing in legacy systems that might be unstructured or haphazardly structured. This has become particularly evident in the last three years, where organizations have fully embraced the use of Intelligent Data Extraction to get precious insights from heterogeneous data sources.

4. Challenges in Data Extraction from Heterogeneous Databases

Along with ever-increasing capabilities of process automation tools, a growing requirement for the extraction and aggregation of structured content from heterogeneous data sources emerges. Financial and commercial organizations hold a substantial quantity of crucial data in conventional relational databases that is routinely read and processed. Price checks, cost monitoring, and competitor analysis are use cases that have traditionally driven merchant data extraction from various sources. Other frequent business logic uses cases require data aggregation from multiple data sources to draw conclusions about potential business opportunities.

Data quality is an important issue in almost every data extraction application. As a consequence, companies that automate stock trading or manage a supply chain risk making important mistakes, unless they monitor the health of their internal bots. Streaming data from target data sources may not be a reliable solution to these problems. Transactional data from relational databases may not suffer from the above-mentioned issues, but they are not the only data sources that organizations access. In order to deliver reliable data extraction, the data quality has to be improved. Integrating the extracted data has to be uniform across any extracted data source. We mean that, when implementing a specific use case, the extracted data from each source must be processed, allocated into different categories, and integrated uniformly into existing operations, such as a machine learning algorithm, existing databases or reporting tools. Ensuring consistency in these processes is extremely important although it is not always feasible to achieve in one extraction operation. Data integration processes across data sources usually come at the cost of complex operations on the delivery side.

4.1 Data Quality Issues

Various data quality problems negatively impact the ability of robots to yield quality data for analysis. Removal of anomalies in the data improves its trustworthiness and can be efficiently handled using RPA robots. Problems such as duplicates, irrelevancy, incoherence, noise, inconsistencies, and lack of adherence to formats are prevalent but reducing them significantly through the RPA process is possible.

Quality also depends on meeting standards of consistency with respect to the organization, but data can also have field mismatches. The standard practices differentiate how data have to be imputed based on category. As an example, a numerical data entry for a categorical attribute is an anomaly while a postal code without a hyphen cannot be a field format-matching standard against which to check. Value range adherence cannot be enforced and an example is letting zip codes with missing characters from the corresponding city or local area. Tabular data is an example of this challenge where unstringed tabular data cannot be fully tabularized by a robot. Faking the tabulation of data spoils the metadata regarding the original data representation.

The task of RPA aims to provide as much quality to the data analysis performed on the final output of the robots working on the data pipeline as feasible. Hence, robots could be used to increase the number of clean, enriched, and consistent entries in the target databases for the entities being maintained. Data distribution performance is vital and can be hampered by decision-making policies as well the deployment and provisioning performance of robots. Data quantity also matters and the massively increasing corpus of worldwide data necessitates this quantity amelioration more than ever.

4.2 Integration Challenges

Relational databases are logically described by schemas and actually stored with a small, fixed number of data formats. On the other hand, most unstructured and semi structured data sources are not equipped with any schema description and their text and markup files are composed, only for their own purpose, with a virtually unlimited variety of textures and base formats. Hence, different sources associated with the same domain may generate data formatted according to heterogeneous structures. This is typical of semi-manual generated data, which have been crafted to satisfy different purposes. This issue does not arise with relational databases.

4.3 Data Format Variability

A further aspect related to the issue of heterogeneity concerns the variability of data formats. Even when they concern the same kind of data set and have the same meaning, they can be expressed in a different way or can have different structures or properties. For example, a specific product could be described in response, in a semantically annotated page, in a product listing on a generic site, or in a review site page, according to different relations. While the first three resources explain what exactly that product is, the last one provides an opinion about it. Data extraction from any of these resource types must deal with variability of data formats and data descriptions. More specifically, data extraction could range from simple scraping solutions completely hardcoded to the resource site being considered, to metadata extraction whose rules are adapted to all the different resource site types, according to some underlying format description model. In between these two levels, we propose several levels of adaptation of the data extraction methods.

5. Technologies Supporting RPA in Data Extraction

Before discussing the integration of intelligent data extraction techniques with RPA, we shed some light on the key technologies that support the intelligent data extraction domain. Intelligent data extraction requires a plethora of techniques, but the major role is played by technology branches such as: machine learning, and particularly deep learning, natural language processing, and optical character recognition. Below we review these three key technologies enabling their role in the intelligent data extraction domain.

5.1. Machine Learning Techniques

In an attempt to increase the level of automation with regards to the machine undertaking tasks without an explicit programming, and to minimize the human intervention, researchers investigated and contributed to the area of general intelligent systems known as machine learning. The algorithms that fall into this area have proven their ability in addressing a wide range of supervised, semi-supervised, or even unsupervised real-life applications. A kind of a new breed of algorithms are the deep learning algorithms, that create a hierarchy of feature extraction during their learning processes. The hierarchical nature and the ability to deal directly with raw input have turned models very useful in data extraction from diverse unstructured or structured data at massive scale and complicated high-dimensionality.

5.2. Natural Language Processing

Natural language processing plays a fundamental role when addressing text embedded in an image or document. NLP is a large and interdisciplinary area concerned with the operations involving written languages with the aim to comprehend, generate, transform, or borrow elements across the languages. From their initial days focused on symbolic techniques including thesauri and grammar, going through machine learning-based approaches using classification, NLP approaches have advanced and nowadays resort to complex deep learning architectures. Last years' significant advances in processing and understanding also very long text sequences has allowed for the development of a new breed of algorithms facilitating the execution of typical tasks with great generalizability to very different applications and higher effectiveness.

5.3 Machine Learning Techniques

Machine Learning (ML) is a domain of Artificial Intelligence (AI) that has achieved considerable progress in recent years. Since most modern ML techniques require retraining with a new dataset to solve a new problem, we will use the term "narrow ML" to refer to the class of Machine Learning techniques that are supported by pretrained models that are either retrained using few examples or not retrained. The main advantage of using narrow ML models is that they require less task-specific data to optimize them for a specific problem. Currently, the vast majority of popular narrow ML techniques are Deep Learning (DL) models that are trained on large amounts of data. As a result, even if narrow ML models are optimized for a specific task, they usually have a larger computational cost than a manually executed logical procedure that solves the same problem.

Due to their training on large amounts of data, narrow DL models can achieve state-of-the-art performance on several NLP tasks, such as named entity recognition, semantic role labeling, part-of-speech tagging, sentiment analysis, machine translation, language modeling, document retrieval, question answering, and speech-to-text, as well as in Computer Vision tasks, such as image classification, object detection, instance segmentation, and semantic segmentation. ML applied to Data Extraction can predict tags or templates that correspond to the format and the meaning of the document data or the model-specific parameters used in logical procedures.

5.4 Natural Language Processing

There are notable, particular Natural Language Processing (NLP) tasks that typically occur during document data extraction: Several machine learning methods may be involved; they are namely named entity recognition, named entity linking, and translation. Named entity recognition is the task of detecting a specific subset of expressions in a text. Named entity linking addresses the task of detecting a set of expressions in a text and linking them with the corresponding, correct entities in a knowledge base. Data-crunching, tree and graph based, template driven, and semantic type based are the four broad categories of methods for named entity recognition. The physical world is very dense and intricate by itself, particularly compared to the space of its abstract representations. Especially, the physical world is dense and intricate compared to its global representations in knowledge bases. Knowledge bases typically offer only fragmentary representations, missing daily events and simple objects. Their coverage is sporadic and their accuracy on other topics is low. Thus, named entity linking from a knowledge base will very likely solve the problem of linking an incorrect name to the wrong entity. Errors in unstructured data may also reflect the bias and the unguided nature of found data corpora. Named entity linking by linking to concepts in knowledge bases includes different tasks. These tasks can include disambiguation of entities, temporal approximation of time-dependent entities, adding attributes to the linked entities, label-selection that is the task of choosing the name by which the entity should be called, and temporal extensibility over the linked entities, extending the validity of the linkage to for example past or future dates. Machine translation, in turn, is the task of transforming an expression in a natural language someone is not familiar with into the corresponding expression in a natural language the person is familiar with.

5.5 Optical Character Recognition

Robotic process automation (RPA), which originally focused on GUI automation, has evolved to address more complex tasks encompassing data extraction, data manipulation, and business logic. Intelligent data extraction from heterogeneous data sources requires a wide array of supporting technologies, including natural language processing, optical character recognition, machine learning, and deep learning techniques. In this chapter, we present a brief summary and discussion of the supporting technologies enabling RPA solutions with a

special emphasis on data extraction and then summarize some of the more prominent RPA products available on the market. The primary purpose of optical character recognition is to support reading printed and/or handwritten text and converting it into a digital format. Most OCR packages today also support forms processing. The primary use case for OCR is reading documents that contain scanned text image information, hence cannot be referred to directly by a called interface, and that require data extraction without the related knowledge bases called for by natural language and rule-based processing techniques. The big advantage of OCR over human and rule-based processing is speed. While it can be trained to read with human-level recognition and accuracy, it generally processes text above a certain resolution at a speed of several million words per hour. Further, rules mapping the extracted text to output variables for a particular document type can be built in seconds.

Consequently, for the past 30 years, OCR has been the mainstay of virtually all production data entry applications. The primary deficiencies of OCR have been the need for training at the character level to achieve resolution, font, and language insensitivity, and the expense and speed of the level of human involvement required to generate reliable character training.

6. Framework for Optimizing RPA

6.1 Assessment of Current Processes

The purpose of a digital transformation initiative is to design target future state processes and to implement a roadmap designed to accelerate that transformation while managing the change on the path to that future state. Before designing that target future state, current processes must be assessed and documented. Without a proper understanding of current processes, existing and proposed automation programs could be rushed without consideration for the long-term return on investment. The best strategy is not to think of RPA as the solution for some business process inefficiencies but to assess existing opportunities for improvement first and then to invest technology resources in RPA only after competent resources identify RPA as the best option for automation.

Traditional iterative process improvement tools can be leveraged to assess existing processes as chain models from process mining. Then, many initiatives such as lean six sigma could be used to document limitations, identify process flows, input and output data, control points, and stakeholders, and measure the basic KPIVs: time, cost, quality, and complexity.

6.2 Defining Optimization Goals

Optimizing existing manual processes is done with a target set of KPIVs and KPQs in mind. Likewise, RPA implementation must target KPIVs. There are often trade-offs made when optimizing existing processes. For example, it is often easier to reduce process completion time if one sacrifices quality. In that case, moving the remaining scrap from a manufacturing / service cycle to the second half of the cycle after automation in some days of a year reduces the current scrap rate and makes the KPIV of origination acceptable. But this is not the objective of automating processes. Internally, individuals creating automating scripts must be employees that are capable of and dedicated to overseeing the quality of the KPQ of the automated RPA process. RPA must improve the total process as a system and not one of its internal steps.

6.3 Assessment of Current Processes

Assisting and automating mind-demanding tasks in advanced processes, such as decision support, data validation, curation, or classification, is the newest development in intelligent process automation. In this context, process optimization enables more efficient implementation of these challenging and resource-consuming tasks. Some innovative technologies, especially Intelligent Data Extraction methods, boost the ease of integrating these technologies in process optimization or time reduction tailored to RPA-integrated processes.

In general, the RPA process assessment is based on an in-depth analysis of the current processes. It outlines the steps in detail to discover unacceptable performance or error rates in time- or resource-demanding activities, enables specifying the requirements of changing the performance and enhancement suggestion evaluation, effort and benefit estimates. The result of the process assessment is a standardized description and structure of processes that is based on the documentation of the processes essential building block-analysis. It subsequently delivers the information basis for discussing, selecting, and preparing activities for automation. This entails deciding on the automation timing and considering the personnel affected by the procedural transformation in the design.

6.4 Defining Optimization Goals

In this section we elaborate on the final phases of optimization of data extraction pipelines, this time focusing on optimization of an RPA process involving heterogeneous databases. Although we will eventually operationalize RPA through the definition of a pipeline involving the deployment of appropriate RPA scripts, we can take advantage of the existence of already-implemented RPA scripts for these pilot tasks.

Framework-based approaches, like the one adopted here, usually follow a hierarchical structure in which the automation job implemented by the framework is the highest abstraction and operationalizes lower-level subtasks defined by a sequence of strategies. Following this view, the general task of data extraction from the target heterogeneous database is defined at the first hierarchical level of the framework.

6.5 Implementation Strategies

Implementation strategies define the best way of carrying out or executing an optimization project. It is vital to point out that not all RPA optimizations need to happen during their production phase and that there are many different strategies for implementing changes. These can be based on several matrices, such as time, benefits, costs, complexity, and flexibility, among others. For example, the current production of the bot could highly affect the results' urgency, offering an

easy choice to be taken, or how many processes are dependent on the bot under analysis could help define the impacts' monetization speed.

In the case we are studying, those strategies are very important for deciding how the Intelligent Data Extraction Cluster should operate and how the bots responsible for the data transfer should be kept available. The automation can be implemented directly through changes made to the scraping bots. For example, by incorporating functionality for revisiting the old URLs of databases currently being pointed. Another option is to simply take the scraping bots out of the production circuit for a predetermined period and leave the URLs indicated unchanged, forcing them to resume activity automatically after a certain period and for certain URLs. In this case, the data transfer is done via back databases, as recommended for future approaches and optimization schemes. Finally, all scraping bots can act in full autonomy through a combination of both strategies, remaining, for example, in a stable circuit for the most part of the time regarding core data extraction, revisiting new scraper URL requests periodically.

7. Case Studies of RPA Implementation

Consistent use of RPA within an organization results in tremendous returns on investment. Typical up-front costs for RPA are about $5,000 to $10,000. Using RPA, the company saves over $20,000 in the first year on a per-robot basis. The return on investment is about 800% to 1,000%. Additional saving occurs after the first year. The growth in the number of RPA contacts has grown by an annual average of 55%. The value of RPA contracts has also grown substantially. The typical company solution includes adopting RPA portfolio of 15 to 20 RPA solutions. Development of an RPA portfolio is time intensive as it engages several business leaders from various departments and requires extensive research on each department functions and how to implement RPA. Eighty-two percent of companies rushing to implement RPA abandoned RPA during testing. The real benefits stem from adding intelligence to RPA. Adding machine learning to RPA will engage artificial intelligence, and it results in IntelliRobots.

7.1. Financial Services

One of the earliest adopters of RPA is the financial services sector. Most businesses found it to be an attractive consideration for how processes can be structured to reduce costs and increase efficiency. Since 2008, banks have been streamlining operations and activity. Banks in North America spend $36 and $210 on compliance and risk for every $1 in profit. The use of RPA could reduce that cost as it results in the savings of $30 to $50 billion annually. Bank of America has deployed over 1000 solutions. It states that knowledge workers cannot be fully deployed unless the mind-numbing processes are automated. Another organization also adopted RPA for reconciliations its settlements. These organizations hope that RPA used in the right circumstances will improve turnaround time and reduce costs.

7.2 Healthcare Sector

RPA has been responsible for transforming the healthcare sector. Compared to investments on other IT technologies, the typical time and cost for obtaining, developing, and maintaining RPA technology is minimal. The return on investment would be substantial. Type of tasks for RPA implementation in healthcare include maintenance of patient registries, appointment scheduling, ensuring the accuracy of medical billing, patient follow-up for billing payments, ensuring that the coding for insurance is complete and accurate, and identifying and preventing fraud as well as during audit processes. Human workforce works harder and their concentration depends on their income and the simplicity of the tasks.

7.3 Financial Services

In 2017, Bank of America Corporation implemented an RPA tool to save 300 hours of manual work monthly for its Corporate Treasury operations, which enables 500,000 transfers of customer funds each day. During this deployment, 450,000 hours were saved from 150 processes. In 2018, with approximately 66 processes in production, the corporation saved an additional 70,000 working hours via Bots. In 2019, Bank of America Security and Centralized Security Services were able to save more than 800,000 hours of work via RPA. The corporation predicted that the utilization of Bots would double. In the same year, Bank of America built over 100 Bots, which reduced the number of outsourced resources from 687 in 2017 to 240 in 2018. During the same period, the corporation also saved on salary resources, approximately 1.3 million dollars in 2017.

In 2019, Citigroup launched a program, ending the year with 200 work done, and around 400 new Bots were planned to be implemented. Citigroup completed 24 Bots for Finance and Operations, saving about 789,000 hours in 56 automations as self-service, launching 306 new third-party Bots made in minutes for close tasks, forecasting work schedules, and tax compliance. In 2020, the corporation automated processes and made automations without code to help investigators file alerts 80% faster. The system includes automations that onboard employees internally and externally to manage money movement, including documentation processes.

7.4 Healthcare Sector

The healthcare sector involves some tasks that need to be performed quickly, accurately, and cost-effectively. One study predicted a potential for over 70% of the operating activities that were being performed in hospitals to be automated. Each year millions of claims are submitted to third-party payors, which verify patients' medical services and procedures performed by affiliated hospitals or physicians, both financially by pre-authorizing those services and procedurally in post-authorizing the payments. Hospitals processing those claims typically experience an enormous backlog of pending issues and, as a result, have started employing outsourced services for managing those organizations. The major functions are often separated into different claim-type lines of business according to the effective date. In addition, workers experience a continuous

increase in difficulty since code and procedure changes in healthcare regulations frequently occur, requiring amendments to the processing systems rules. Hospitals have therefore begun to adopt its process improvement initiatives and internal support staff. RPA has also been used for operations in many other areas of healthcare settings.

Top 30 medical groups, including healthcare delivery services, are mostly focusing on RPA for improving processes in coding, scheduling, registration, and revenue cycle management. RPA's virtual assistants and bots can be found in both administrative offices and hyper-connected operating rooms. In administrative offices, bots typically perform mundane tasks, including answering calls, maintaining schedule for billing, making payments, updating records, and reconciling accounts. In hyperconnected operating rooms, RPA is an emerging technology offering safety and error reduction, enabling reliable monitoring of the network's electronic devices while triggering corrective actions in case of detection of any faults. Operations are carried out by virtual robots for high-volume, low-complexity repetitive tasks like clerical work. RPA enables doctors and nurses to spend more time on providing quality care rather than data entry. With increasing data circulation in a hyper-connected operating room – from equipment alerts to scheduling requests – organizations are embracing for RPA.

7.5 Retail Industry

The retail industry abounds in ready opportunities for optimizing RPA platforms. Intelligent data extraction can greatly aid the procurement process, where requests for quotations, price lists, order confirmations, and shipment updates are received from multiple business partners such as suppliers, customers, freight forwarders. These documents are typically received via emails, sent in multiple formats and structures, and sometimes written in different natural languages, hampering the automation and increasing operational costs. The data received is then manually validated, formatted, and inserted into specialized business applications to support procurement operations. A well-tuned RPA platform can automate this document-processing stage, creating the possibility for an end-to-end solution.

8. Best Practices for RPA Optimization

When RPA is deployed and maintained, it will generate positive impacts on its implementation and make it easier to address future automation initiatives. Understanding how to make RPA successful over the long term is key for maximizing ROI and generating continuous value.

Continuous Improvement

Robotic process automation is not a plug-and-play technology. Automation provides a chance to rethink and reassess how processes flow, how decisions are made, the tools and systems being used, and the documents that are being shared. Short cycles of testing and reconfiguring create improvements much faster than any set of written processes can convey.

Stakeholder Engagement

RPA is a change management initiative that involves people, process, and technology. Above all, this is about people. RPA takes away the mundane tasks and enables workers to spend their time where it matters most. Identifying stakeholders and getting them involved in every stage of the RPA journey is crucial. Those directly engaged in the processes must provide input on how they do their jobs, their pain points, and what could feasibly guide automation. They must also be available during testing and help identify the optimal way for bots to process transactions.

Performance Metrics

Only a small percentage of executives say they are maximizing management and operations. KPIs should measure the effectiveness of RPA as a governance strategy but also the processes being automated in order to identify further areas for optimization. When deciding what to measure, scalability, benchmarks, business impact, impact on customer experience, and impact on employee experience should all be considered. Establishing these metrics at the start of the RPA journey – and being transparent about them – will improve the chances of success.

8.1 Continuous Improvement

As with any other technology, RPA should not be considered a "faça tudo" solution that once implemented is no longer required to be monitored and evaluated for improvements. An RPA deployment should be continuously improved, kept up to date, and evaluated for new process improvements. Nearly 90% of enterprises are presently using or investigating RPA adoption and implementation, whereas more than half of organizations have adopted or are in the process of adopting hyper automation, which is amplifying and extending the use of Automation technologies, including RPA. At the same time, these RPA solution adopters may want to eventually retire some of their early adopters RPA efforts as evidence that Cloud Native Platform services, AI, and Machine learning may be increasingly able to address the processes that were originally implemented with RPA. Taking together, these facts imply that an organization might be able to leverage an increasing pile of potential automation capabilities that can be stitched together effectively over time. It also implies the recommendation to not just put RPA efforts "on autopilot" after the first release of automation or after the initial burst of automation. While there might be a natural drop-off of interest initially in RPA, organizations should be leveraging a Center of Excellence and other models for continuous improvement.

Continuous improvement related to RPA implementation includes the fact that using RPA technology does lend itself to periodic review of which of the processes are currently automated, as well as which ones were previously reviewed for implementation with RPA but not acted upon. These periodic evaluations of the set of implemented automation RPA solutions assisted and supplemented by a well-defined repository of the set of implemented and currently staged RPA automations, not just for discovery purposes in order to get visibility on documentation, but also to facilitate the management of the automations and their evolution over time.

8.2 Stakeholder Engagement

Stakeholder engagement is an essential area of concern in RPA system optimization. In this chapter, we review the group of stakeholders typically identified, including their responsibilities involved in each phase of the RPA lifecycle. RPA tools are specifically designed to communicate with software user interfaces, automating manual processes. Traditionally, those business processes are defined by business analysts and performed by business professionals. RPA robots relieve employees on tedious activities that require high levels of concentration and attention to detail. However, RPA robots are not responsible for the business processes automation from end to end, only the execution of the repetitive rule-based task. In this way, RPA optimization concerns not only the robot operations but also business-related tasks from which they rely on. RPA robots make intensive use of business applications and databases. Therefore, in addition to the robot logs, support from IT employees can assure improvement on performance and can resolve potential conflicts of resource sharing. Finally, it is also important that senior leadership sponsors business units' RPA initiatives, assuring fund allocation and resources engagement for the RPA program.

The RPA lifecycle maps naturally to both the RPA optimization phases and the RPA tools infrastructure. RPA tooling operations can guide actions and allocate responsibilities among the stakeholders. For example, task discovery occurs before the RPA know-how sharing phase. The RPA tool can suggest candidate processes for automation based on usage logs. RPA suitability analysis can be performed by business analysts or domain experts.

8.3 Performance Metrics

The performance of the RPA solution requires effective measurement and constant monitoring, but selection of the right metrics is not always obvious. Many organizations rely on service level agreements, which effectively measure response time of the automation in delivering business outcomes or intermediate milestones. Other organizations actively measure the change in labor allocation, either tracking hours billed to the specific function or hours recorded internally. Such charge code tracking assumes workers are diligently assigning time to the appropriate codes in time entry applications, which may not always be the case leading to data integrity issues. This metric further assumes that an appropriate process definition exists to isolate the effects of automation. Other performance indicators measure relatively less dynamic indicators. These could include the percentage of paperwork rejected or the percentage of paperwork requiring exception handling by a human at some point due to incorrect data extraction.

9. Future Trends in RPA and Data Extraction

A nascent trend is emerging whereby vendors promote a convergence of different technologies intended to generate a greater level of automation in mundane tasks otherwise requiring human intervention. RPA vendors also provide some basic AI capabilities as first-line solutions for higher levels of automation. The result is the growth of the combined use of RPA or intelligent RPA with AI-based technologies, such as computer vision, process and task mining, natural language processing, and machine learning, generally as modular components that enhance RPA solutions. So far, intelligent data extraction has been one of the major RPA use cases. The advent of cognitive services by cloud vendors and the additional emergence of specialized vendors for intelligent data extraction from documents enhance RPA's capabilities for the purpose of enabling better solutions.

9.1 AI and Automation Convergence

Areas like document understanding, entity recognition, sentiment analysis, and chatbot development are increasingly merging NLP research with the real-world needs of businesses. Market pressure to get new products features to market fast is changing how RPA developers think about documents. Dev teams may be drawn into developing a 10-year strategy to use massive amounts of labeled data, vast compute power, complex modeling, and big distributed inference, thus wedging burdensome, narrowly focused diffusion, transformer, or GAN-like models into their future pipeline. Nevertheless, many also realize that the functionality offered today can sometimes provide valuable, immediate assistance to short-term projects. Toolkits can be integrated into RPA processes to help provide time-sensitive solutions with acceptable levels of automation. Using advanced, cloud-based labeling services, these creators can easily upload unlabeled documents. Then, hugely powerful pipelines can label thousands of documents in a few days. NLP pipelines and chatbots can provide tight integration with LLM services, allowing developers to provide LLM-based augmentation within existing applications.

Despite the pressure to integrate and leverage advanced toolsets, Robotic Process Automation (RPA) provider plans still contain an underlying philosophy: document understanding or advanced data extraction is the norm. Today's advanced pipeline/engine-based services have become so powerful that for some companies, it doesn't make sense to do everything in-house, let alone architect a strategy that focuses on RPA. Perhaps this is why RPA product strategy is shifting; data extraction for RPA is changing. Understanding the intersection between data extraction and RPA development is paramount for RPA developers. The RPA developer's world is shrinking. Cloud-based data extraction services have appeared, offering numerous highly specific, one-off extraction schedule pricing services and commercially viable unclaimed template option services.

9.2 Enhanced Data Security Measures

Data security is always a matter of concern when an automation solution interacts with sensitive data such as customer information or intelligence data. Some depositors were worried that the bank would not be able to protect their data from hackers. Once automation is in place, sensitive documents may be floating everywhere without supervision since robots can access any document at any time. So how do we enhance data security in data extraction?

For sensitive data-extraction activities, an extraction bot can combine with blockchain technology to make it more secure.

Essentially, every sensitive data extraction will generate a hash file that contains the fingerprint of the transaction for forthcoming reference. A document confirming delivery is scanned when a delivery is made, and the corresponding hash file is generated. Any modification of the document will be checked against the previous hash to ensure its validity. Such an approach defeats any human desire to modify documents for discreditable gain. However, implementing blockchain technology may add extra overhead in terms of implementing non-repudiation and longer transaction verification time. Data validation will be needed to run separately, or it may degrade the overall execution speed.

10. Ethical Considerations in RPA

As RPA can be leveraged to fulfill complex tasks largely without human involvement, by extending programming possibilities to non-experts, it raises ethical concerns particularly in view of potential increased bias, risk of discrimination, or conversion of workforce for unfair interests, as well as induced or augmented inequality. In this section, special emphasis is put on the challenges of robotic process automation from a global perspective regarding data privacy concerns and bias in automated systems.

Data privacy protection is an essential part of safeguarding rights and freedoms of natural persons. Even if data are only temporarily processed for the fulfillment of a task, in the case of RPA personal data could be retrieved, used and made accessible to robots, and, then, actually also to robots' owners or operators. Ethical data patterns must follow a clear "human-centric" approach that ultimately prioritizes the well-being of those concerned against the interests of their owners or operators, and should create crystals of trust at ecosystem level.

10.1 Data Privacy Concerns

Data is key for a company, as one of their most valuable assets are their customers. Businesses are beneficiaries of customer data for personalized marketing programs, like discounts and offers, in order to increase sales volume and consumer loyalty. But, at the same time, customers become victims of data privacy breaches related to data extraction, use and sharing. There are a lot of companies today that buy, make use and share consumer data without personal consent concern. The increased number of cybercriminals aiming to utilize these data flows is impacting the consumer daily. The right to privacy remains a constant battle exercised through the courts by consumers. And technological solutions, like prevention and protection measures, are being developed, however, they are still not enough. In this dominating world of cybernetics, technological solutions should be ethically used to protect personal information.

10.2. Bias in Automated Systems

Although AI systems can lead to more fair outcomes than traditional approaches, the decision-making models that AI systems use can be biased based on data, by inducing inequality and discrimination, or at the algorithmic level itself when such design is not accounted for. One of the models that led – and leads – more to ethical issues in AI is the neural networks, which given Input A and Input B, provide output C. When, for instance, Input A is a person's face and Input B is the probability that he or she is a criminal, a natural question arises: why is a black person more likely to be classified as a criminal than a white person? After all, the neural network just learned from the data provided to it that this person was a criminal, but it is inscrutable to understand the criteria that the model has taken into account to provide classification C – whether comparative skin tone, cheekbone configuration, or any other detail that relates to race. Similar questions can be raised with respect to sex and to any other feature that characterizes a set of people when another set in a society or in a comparable group has not been used to train the model, is not proportional in statistical sense to the considered set, or is simply not specified. Such concern is not minor; AI decision-making, for instance, in cyber-crime risk assessment relates to a business market of U$ 600 billion per year! In banking, risk classification prevented U$ 100 billion from being lost, but it can cause much damage when a wrong or biased decision is taken. Similar consideration can be done with respect to AI investment related to customer support.

11. Conclusion

The growing emphasis on automating the capture of data from a variety of electronic documents is being driven forward by the realization that a high level of data extraction automation can lead to significant efficiency savings in higher education admissions offices. Rapid growth in the number of datasets being added to a heterogeneous database of datasets means that it is becoming increasingly cumbersome for potential data users to independently identify the appropriate and necessary information to access a specific dataset. An RPA bot educated with sufficient knowledge to identify and extract the required pair of data items associated with each new dataset from a collection of manually compiled datasets would expedite the completion of this repetitive work, enabling an individual's limited amount of available time for engaging with a potentially valuable data resource to be spent on data interrogation instead.

This work has described how the capabilities of RPA technology can be optimized to best enable the intelligent automation of this extraction of pairs of data items from heterogeneous data sources. By conducting a small suite of preliminary exploratory studies that systematically evaluated available commercial RPA products for their suitability for this Intelligent Data Extraction process, as well as simultaneously optimizing achievable performance levels, the conclusions presented here should act as an informed guide for any practical implementations of RPA-based workflows. In summary, the results described here highlight the importance of managing and optimizing each, RPA tool sequencing, IDE for specific use-case customization, enabling human-centered workflows, and runtime efficiency optimization, and appropriately educating the RPA bot for the provisioning of suitable task-based demonstrations, whether relying on single or multiple demonstration examples, as a method of preparing the RPA bot for being able to achieve the accurate completion of data extraction for a wider range of real-world use-case examples.

References

[1] Galkin, M., Mouromtsev, D., & Auer, S. (2015). Identifying Web Tables - Supporting a Neglected Type of Content on the Web.
[2] Karpathiotakis, M., Sérgio De Oliveira Branco, M., Alagiannis, I., & Ailamaki, A. (2014). Adaptive Query Processing on RAW Data.
[3] Rizk, Y., Isahagian, V., Boag, S., Khazaeni, Y., Unuvar, M., Muthusamy, V., & Khalaf, R. (2020). A Conversational Digital Assistant for Intelligent Process Automation.
[4] Kumar, A. & Ré, C. (2011). Probabilistic Management of OCR Data using an RDBMS.
[5] Bernstein, V. & Afanassenkov, A. (2020). Unsupervised Data Extraction from Computer-generated Documents with Single Line Formatting.