



Nuix in the Cloud

Automating data processing in Azure or AWS

Driving Change Whitepaper Series

December 8, 2021

Disclaimer

This document is provided “as-is.” Information and views expressed in this document may change without notice. This document does not provide any legal rights to any intellectual property in any Rampiva product. This material does not represent a commitment from Rampiva to provide software or services to the recipient. Recipient may copy and use this document for internal, reference purposes.

Please do not distribute without written permission from Rampiva.

© Rampiva Global LLC, 2021 All Rights Reserved

Table of contents

1.0	Executive Summary	5
	Rampiva Key Benefits	5
2.0	Introduction to Rampiva	6
3.0	A Client's Considerations	7
	3.1 Optimal Azure or AWS Infrastructure	7
	3.1.1 No Idle Time - Spin up and Down Compute Resources as Needed	7
	3.1.2 Ability to Use Price-Optimized Spot VMs	7
	3.1.3 Ability to Distribute Single Job Across Multiple Nodes	7
	3.1.4 Example Azure or AWS-First Architecture Diagram	8
	3.2 Cost Management	8
	3.2.1 Rightsizing Azure or AWS Resources	8
	3.2.2 Machine Spin Up and Spin Down	8
	3.2.3 License Assignment	9
	3.2.4 Direct Data Upload	9
	3.2.5 Self-Service	10
	3.2.6 Deduplication	10
	3.2.7 Direct-to-Relativity Upload	10
	3.2.8 Real-Time Performance Monitoring	11
	3.3 Optimized Performance	12
	3.3.1 24-Hour Job Scheduling Queue	12
	3.3.2 Workflow Automation	12
	3.3.3 License Management and Rightsizing	13
	3.3.4 Circumventing Workstation	13
	3.3.5 Deduplication	13
	3.3.6 Data Segmentation	13
	3.3.7 Computer-Assisted Quality Assurance	14
	3.3.8 Notifications	14

3.4 Monitoring	14
3.4.1 Job Queue	14
3.4.2 Notifications	15
3.4.3 Dashboards	15
3.4.4 User Access Controls	15
3.5 Automation	15
3.5.1 Workflow	15
3.5.2 Scheduler Job Queue	16
3.5.3 Centralized Reporting	16
3.6 Agility	16
3.6.1 Fully Isolated Template-Based Deployments	16
3.6.2 Fully Isolated Playbook-Based Deployments	16
3.6.3 Centrally Managed Deployment with Isolated Processing	17

1.0. Executive Summary

A client is expanding service and support capabilities in European countries with strict privacy requirements that can no longer be delivered by resources based in post-Brexit United Kingdom. To enable clients in these regions, a client is pursuing a cloud-first model, leveraging Azure or AWS resources for scalability, capacity pricing, and simplified administration.

However, data processing in Azure or AWS comes with risks and restrictions, including:

- Staff costs of manual administration;
- Risk of resource “leakage” when machines are left running past task completion; and
- Performance relative to on-premises hardware with comparable specifications.

To help address these challenges, the client’s has asked that Rampiva provide details on how our Automate platform manages Nuix resources, Workflows, and Cases in Azure or AWS. After a general introduction to the Rampiva Automate platform, this document will speak to six areas:

1. Optimal architecture in Azure or AWS;
2. Azure or AWS cost management;
3. Optimizing performance with Azure or AWS hardware;
4. Monitoring tools;
5. Automation of Azure or AWS and Nuix resources; and
6. Agility for expansion into new regions.

By evaluating these six areas, the client expects to determine whether Rampiva Automate can provide sufficient value to be included as a near-term partner while spinning up new environments in Europe.

Rampiva Key Benefits

Rampiva delivers savings from reducing Azure or AWS utilization and reducing staff-time setting up processing jobs, while improving the performance of Nuix Workers, the productivity of the overall environment, and the ability of a client’s team to monitor, manage, and support the environment.

2.0. Introduction to Rampiva

Rampiva Global LLC is a multi-national software development company focused on making quality data forensics and analytics widely available. Founded by Daniel Boteanu in 2017, Rampiva has focused on automating and administering the Nuix data processing engine. Rampiva was the first firm certified by Nuix as “Master Developer” in 2020, supports 25+ active clients in 7 countries, and runs over 1,100 Nuix Workers with our core Rampiva Automate product. Three Rampiva clients have more than 100 Nuix Workers in their environments, processing more than one petabyte through Nuix+Rampiva annually.

Rampiva Automate delivers resource and Case automation functionality that enables architecture management, the execution of tasks in the Nuix Case, and real-time reporting across all projects. To deliver this capability, Rampiva Automate has three layers:

- **Workflow:** A thick-client program that leverages the Nuix Engine's Java API to automatically execute a defined series of processing, search, transformation, and reporting tasks.
- **Scheduler:** A web-based program that manages Nuix license, hardware resources, and jobs based on a prioritized queue and the selected Workflow.

Scheduler allows for granular user-defined role-based access controls, with the user authentication being performed using Microsoft Azure or AWS AD or other industry-standard single-sign on protocols.

- **Dashboards:** A centralized data repository that tracks performance and utilization across all jobs and Nuix Cases that is accessible by BI tools such as Power BI via an OData feed or native SQL.

This platform provides teams with access to a 24-hour processing cycle that eliminates Nuix Worker idle time, reduces the staff-time required to stage and process Nuix Cases, improves user access controls, and allows for real time monitoring, notification, and reporting.

Rampiva has added a new layer to our ability. We are able to process more data, quicker and with less human error than before. Our technicians can focus on more pressing client issues and leave the processing and reporting to Rampiva. No matter how complex the workflow, Rampiva can handle this with a single, or multiple workflows if necessary.

**Associate Director, Multinational
Discovery Services Company**



3.0. Considerations

The client has provided six areas of inquiry, focused on how Rampiva Automate can create a more efficient and productive system for running Nuix in Azure or AWS. For each area, Rampiva has provided specific features and capabilities that improve control, reduce waste,

accelerate performance, increase productivity, and provide administrative insight and control.

Where available, Rampiva provides real-world impact metrics from active clients to highlight the potential benefit of this partnership.

3.1. Optimal Azure or AWS Infrastructure

Rampiva Automate has native support for Azure or AWS and has clients actively processing data with Nuix hosted in an Azure or AWS instance.

As part of this architecture, Rampiva has:

3.1.1. No Idle Time - Spin Up and Down Compute Resources as Needed

A key piece of the Rampiva Automate system – the Scheduler Job Queue – manages the assignment of Nuix Workers and hardware resources from completed jobs to new jobs. For clients operating in Azure or AWS, Rampiva will also spin down the machines that were running the finished job. This means there is little-to-no time that the machines are up (and incurring charges) but not actively working on some aspect of the Workflow.

This is discussed in more detail in Sections 3.2.2, 3.3.1, and 3.4.1.

3.1.2. Ability to Use Price-Optimized Spot VMs

Rampiva Automate can take advantage of Spot VMs which have a highly discounted pricing (up to 90%, according to Microsoft). Typically, Spot VMs would not be appropriate to use for processing data with Nuix, because an interruption of the Spot VM would interrupt the job in progress and risks corrupting the Nuix Case. However, with Rampiva Automate, the Job and the Nuix Case are managed by a lightweight on-demand VM, and remote workers run on Spot VMs. In case of an interruption of the Spot VMs, the Job continues to run, and the integrity of the Nuix Case is not impacted.

3.1.3. Ability to Distribute Single Job Across Multiple Nodes

The on-demand VM hosting the Nuix Case and managing the overall date-load process can be a relatively lightweight machine. This helps manage the interrupts in Spot VMs and stops high-cost processing resources during the Nuix operations that do not require significant machine resources (such as search and tag, filtering, reporting, item set creation, and more).

3.1.4. Example Azure or AWS-First Architecture Diagram

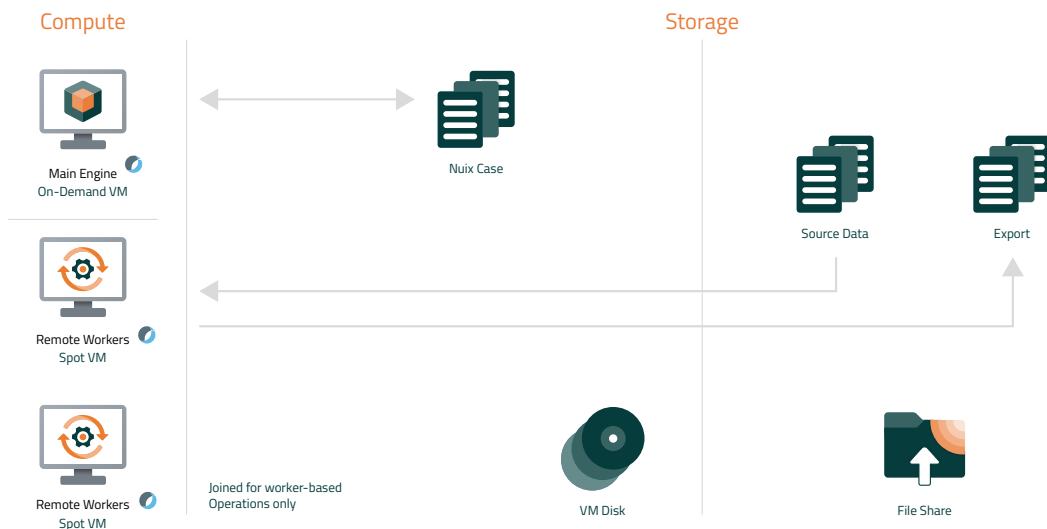


Figure 1: Sample Cloud-First Azure or AWS Architecture Diagram

3.2. Cost Management

To help clients manage the cost of processing data in Azure or AWS, Rampiva Automate provides control of the environment, accelerates the data processing itself, and provides real-time insight into current and past performance.

3.2.1. Rightsizing Azure or AWS Resources

Azure or AWS hardware resources are highly customizable, allowing users to design exactly the right machine for their project. This is a big value-add for data processing teams, because projects, timelines and budgets are all variable. Aligning the right hardware resources with a project can save time – and, in Azure or AWS, money.

Rampiva Automate’s Resource Pool functionality makes it simple for users to administer this activity. Administrators can define a portfolio of machines with various resource specifications and save them for use later. Sophisticated teams can go one step further and design Resource Pools that are only accessible by certain analysts – or individual clients.

Make it easy for your team to select the right resources for the job.

3.2.2. Machine Spin Up and Spin Down

The most resource-intensive steps of processing data in Nuix are the four operations that transform the data – loading data, OCRing files, imaging pages, and exporting data. These are also the most expensive steps when running Nuix in an Azure or AWS instance, because good performance requires well-resourced machines. It is important to ensure that these machines are only used for these tasks, and that they are spun down as quickly as possible.

There are three ways that Rampiva Automate addresses this need:

- 1) VMs are only powered on when there are jobs in the queue that need to be worked on. As soon as the job is complete, the VM is automatically powered down using a robust mechanism that accounts for potential errors in the workflow, or unexpected crashes of the Nuix Engine or of the underlying Java Virtual Machine (JVM).
- 2) When performing non-worker-based tasks, such as searching, tagging, running scripts or automated QC steps, only the main VM is powered on. When performing worker-based tasks such as loading data, running OCR or export, additional spot VMs are started to offload the compute-intensive workload to price-optimized resources.
- 3) Improvement in the overall time it takes to get data through Nuix, which is achieved by chaining all operations in the workflow and thus reducing idle time, reducing staff time, and reducing or automating QC checks;

Together, this approach reduces the cost of running Nuix in Azure or AWS by maximizing the productivity of uptime.

Benchmarking: Manual Nuix processing teams routinely show 20%+ idle time, meaning Nuix Workers are checked out from the pool but are not processing data. Assuming an hourly Azure or AWS server rate of \$8.75 and 6 servers operating 24 hours per day, a comparable idle time would add \$252 in unnecessary spend per day, or \$91,980 per year.

3.2.3. License Assignment

Assigning Nuix Worker licenses to the next queued job – and, returning licenses to the pool when the job is completed – maximizes the availability of Nuix resources and the number of productive hours in a day.

Increased productivity is valuable for two reasons. First, the more data a team can process per day, the more money that team can make from processing data. Second, the faster any individual set of data is processed, the sooner review can start, and the happier the client.

By managing Nuix Worker license assignment, Rampiva Automate ensures a productive environment, a low marginal cost per gigabyte, and achieved SLAs.

3.2.4. Direct Data Upload

Rampiva Automate allows teams to submit data on remote systems to a client's local instance of Rampiva Automate. This reduces the need for 3rd party data-upload products, enables the submitting party to schedule a job that will start once the data upload is completed, and ensures a unified set of data, from submission to export.

Staging data submitted over FTP or on a drive can be time consuming, risks human error, and creates friction in the path from source to processing and review.

3.2.5. Self-Service

Authorized third parties can access Rampiva Automate with varying levels of permissions. By combining this feature with strong User Access Controls, Rampiva users can empower clients to monitor (only) their own jobs on the queue, submit data, load data directly into Rampiva for processing, and even start jobs.

This helps align resources across client schedules, reduces the burden of project management and reporting (particularly for smaller jobs), and helps maximize system utilization.

The Rampiva Automate interface can be customized in some ways, to fit with the clients branding guidelines.

3.2.6. Deduplication

The Nuix engine is the industry-leading tool for processing data—particularly for very large projects.

Their approach allows teams to create—and add data to—databases with hundreds of millions of items. Alternatively, teams can maintain many smaller databases and then compound them together to create a unified environment.

In both scenarios, there is one bottleneck that can be challenging: identifying items that are duplicates across all of the custodians and all of the cases. In an Azure or AWS environment, this approach becomes prohibitively expensive.

However, Rampiva Automate users can take a different approach—one that saves times and ensures repeatability, by leveraging dedicated memory- and compute-optimized data deduplication operations. One client said “this is the difference between 2 weeks and 2 seconds” for accelerating Nuix deduplication.

Benchmarking: When adding data to a 200 GB Case, one client was unable to deduplicate against existing items despite 225 GB of memory allocated to the task – the process would run for days and then crash. By leveraging the Rampiva Automate deduplication feature, the items finished deduplicating in 10 minutes.

3.2.7. Direct-to-Review Upload

An immediate priority for most eDiscovery teams is to get data into review platforms as quickly as possible.

There are three areas where Rampiva Automate helps make that happen:

1. **Proprietary Rampiva Connectors** to Relativity On-Premises, Relativity One, Nuix Discover On-Premises, and Nuix Discover SaaS that allows the automatic uploading of data from Nuix Cases to these review platforms;
2. **Automatic Load Balancing** that splits Nuix Production Sets into smaller sets and parallelizes the uploads, which in turn smooths the export from Nuix and the ingestion into the review platform. This helps to ensure that the first set of items is in the review platform more quickly and that the total time between the start of the export and the end of the ingestion is shorter as well.
3. **Quality Assurance Intelligence** to automatically check for common upload errors, like file-path names that are too long, or items with invalid dates or no content and then, retry the item with adjustments that address the original error, all while logging the patches applied.

Enabling Direct-to-Review Exports, Rampiva Automate accelerates speed to results, reduces the time spent fixing errors, and allows end-to-end automation.

3.2.8. Real-Time Performance Monitoring

In addition to writing the standard Nuix Case, Rampiva Automate will also store over dozens of performance metrics on every job, session and operation run through Rampiva. By connecting this database to a BI platform, using the OData protocol, teams have real-time access

to powerful performance monitoring dashboards. These same metrics can also be pushed into a SQL database or to Matter Management and billing systems for advanced analytics.



Figure 2: Real performance data from a large ediscovery vendor

The immediate benefit is simply the ability to easily observe performance, whether that's throughput, average caseload, or speed-to-review.

However, because this data is real-time and comprehensive, advanced clients can build their own analytics to measure activity by time-of-day, observer performance variation by file-type given different Resource Pools, test the marginal throughput gains of additional Nuix Workers, and more.

These insights can help set client expectations, optimize processing strategies, and plan capacity.

3.3. Optimized Performance

Rampiva has identified five Key Performance Indicators for measuring Nuix productivity: Throughput per Nuix Hour, Activity per Day, Uptime, Staff Time per Case and Speed to Results.

Taken together, these metrics allow a meaningful comparison between departments, or within a department and across periods.

3.3.1. 24-Hour Job Scheduling Queue

The most expensive aspect of a Nuix data processing environment is time. For many clients, commitments and turn-around time is measured in days – sometimes even hours. Rampiva Automate ensures that every team has access to their Nuix environment 24-hours a day. While large data centers have the resources to keep their Nuix workers processing, Rampiva Automate gives that scale to even the smallest instance.

While this will not change the throughput of the Nuix Workers or the Staff Time per Case, it dramatically improves Activity per Day, Uptime, and Speed to Results.

A great way to measure the impact of this feature is Idle Time before First Export. These are very expensive hours, both because of the Nuix Workers that are not being used on the next job, and because of the review hours that cannot be spent yet. It is common for the bulk of these hours to be incurred off-shift between 6PM and 8AM – a load job finishes, but the OCR job cannot start until a (soon to burn out analyst) wakes up at 2AM to sign in through VPN and kick off that next step. The current job is delayed getting to the review platform, and so is the start of the next job in the queue.

Benchmarking: As shown in Figure 2, a large eDiscovery vendor using Rampiva has a 95% Efficiency rate (Active Nuix Hours/Total Nuix Hours) and just 1,220 Idle Hours before First Export in 6 months. That is less than 15 minutes per Case.

3.3.2. Workflow Automation

The second most expensive aspect of a Nuix data processing environment is usually the person setting up and running the processing jobs. One client estimates that it took 90 minutes for one of their analysts to stage, index, OCR, search, tag, create production sets, create a processing report, and exporting the final item set – including the QC work that goes into make sure each stage was executed correctly. There were 88 separate clicks required to complete the processing of the average load file.

After adopting Rampiva, this same department estimates less than 5 minutes per Load File – making their Analysts 18 times more productive.

3.3.3. License Management and Rightsizing

Also, the Nuix Workers themselves are expensive. Rampiva helps optimize this resource by returning worker licenses to the pool for use on other jobs.

Section 3.2.3 describes how this feature helps control cost in an Azure or AWS instance – when jobs are done.

However, clients can build in a little extra efficiency on the Nuix license management side. It is common for clients to check out 4, 8, 16 or more Nuix Workers to execute the resource-heavy tasks (Load, OCR, Imaging, Export) and then hold onto those Workers when executing the lightweight activity like search-and-tag, applying data filters, applying custom metadata, creating production sets, or generating a report.

An advanced Rampiva Workflow can return all but 1 Nuix Worker to the pool when executing these tasks – making them available to start the next processing job.

3.3.4. Circumventing Workstation

A tactical improvement in performance when using Rampiva is that teams do not have to open Cases with the resource intensive Nuix Workstation interface. Designed for fact-finding and forensic triage, Workstation can be a bottleneck for very large Cases.

Although the cases create by Rampiva Automate are 100% compatible with Nuix Workstation, Rampiva Automate is built on top of the Nuix Java API, so it does not open Workstation before executing operations.

This can make it a lot easier for teams to work with large Cases – particularly Compound Cases.

3.3.5. Deduplication

As discussed above, Rampiva Automate accelerates deduplication. This creates cost savings in Azure or AWS and improves the overall performance of the team.

3.3.6. Data Segmentation

Nuix users will often run large jobs using a lot of Nuix Workers – even though the marginal Throughput per Nuix Hour decreases as Worker counts grow. In a manual operating environment, this makes sense – it is easier to run a single one terabyte job rather than segment the data and stage ten 100 GB jobs.

With Rampiva Automate, it becomes routine to split source datasets into manageable chunks, assign an optimal Worker Count, and queue the entire set. This approach helps maximize of the availability” between “maximize” and “of” so it reads “This approach helps maximize the availability of the Nuix Worker itself.

3.3.7. Computer-Assisted Quality Assurance

To maximize the value of automation, it helps to minimize human check-in points. Human intervention creates friction across a variety of fronts – so, Rampiva helps users find ways to remove these by:

- Enabling script-based logic, with complete access to the underlying Nuix Case, that will iterate through likely remediation steps;
- Allowing Analysts to define Parameters that influence conditional Workflow stages during creation; and
- Branching Workflows so data problems that need to be resolved manually do not prevent the rest of the items from passing through to review quickly.

This approach helps drive productivity across the entire platform and focuses expert attention on the problems that really require it.

3.3.8. Notifications

Rampiva Automate connects with email and collaboration platforms such as Microsoft Teams and Slack so that alerts, reports, and notifications can be sent to defined distribution lists or channels. Alerting a client that their job has started, sharing an error report with a project manager, or pushing a completed processing report to the whole team is easy and fast. Advanced applications rely on custom logic, conditional Workflow steps and report design to escalate urgent results as appropriate.

A well-designed Notification plan helps reduce the administrative burden of increasing Caseload.

3.4. Monitoring

Coordinating resources across regions and teams is an essential part of managing a client's data processing requirements. This can be challenging across multiple instances, particularly given that each Nuix Case is a stand-alone database. Rampiva provides several key features that help address this need.

3.4.1. Job Queue

Monitoring the Job Queue addresses scenarios where, for example, one client wants to know when they can expect their data for review, but a huge Case is consuming all of the available Nuix Workers during off-hours. In this case, the Project Manager has to wake up the Analyst who started the large Case to login, check the progress, and estimate on when the resources will be available.

With Rampiva Automate, everyone on the team can see the progress of the big job – and keep an eye on what is coming down the pipe. Some clients have a central Job Queue with regional Resource Pools, and others give their teams cross-regional access permissions.

3.4.2. Notifications

The same notification capability that helps teams administer a growing caseload also ensures improves Case and process monitoring. Leaders can focus on their immediate priorities, with the confidence that they will be alerted when progress is made, or issues are encountered.

3.4.3. Dashboards

Real-time reporting is critical – particularly if users take the step of incorporating Case-specific metrics like Custodian counts, hit reports, and/or Item Date metadata. Rather than using a Nuix Worker to manually open a dozen Cases to check whether a frequent flier has any data from 2014, this information can be directly seen in the Dashboard.

Worried about performance? Check the Dashboard.

Wondering which Analysts have worked on this project? Check the Dashboard.

3.4.4. User Access Controls

Rampiva's User Access Controls allow clients to govern who on their team can access which Workflow Libraries, which Resource Pools, and which client environments. Additionally, users can be given processing rights through Rampiva without that same user having access to the Nuix Case itself. This control reduces the need for monitoring activity because the client can control against resource hogging, user error, and data access.

3.5. Monitoring

Rampiva Automate focuses on three specific areas – the operations executed on data, the management of resources, and centralized reporting. Specifics of this functionality are described previously in this document, and the remainder of this section will focus on a high-level summary.

3.5.1. Workflow

Rampiva Automate allows users to design unique Workflow and store them in Libraries. These Workflows define a specific order for operations to be executed, including activity before the creation of the Nuix Case, data processing, and activity in the Case itself. Rampiva can automate all functionality from Nuix Workstation. By leveraging Custom Scripts as part of their Workflow, users can also automate interaction with systems outside of Nuix. Finally, users can leverage Parameters to build conditionality into the Workflows.

Most Workflows have 20-40 steps, but advanced Workflows can have more than 100 steps.

Users can also define new Workflows for specific jobs using the Rampiva Workflow Wizard, a guided process that takes the user through a series of questions and creates a customized Workflow within minutes, ready to be used, or that can be further customized in the Workflow Designer.

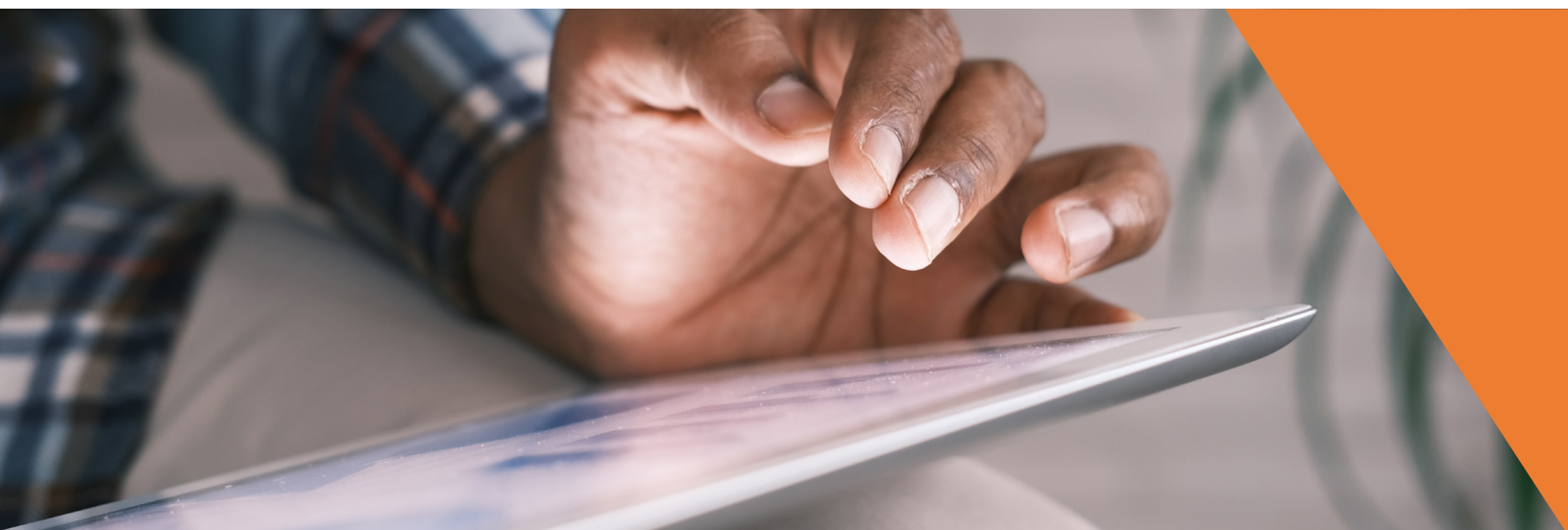
3.5.2. Scheduler Job Queue

Rampiva Automate assigns Nuix Workers and hardware resources to jobs based on their priority in the Scheduler Job Queue – and will recover those resources when the job is completed.

Jobs can be set to automatically run based on external, script-based triggers, made conditional to triggers within Rampiva Automate itself (such as a processing error) or set to recur on a defined timeline. As described in Section 3.1.1, for clients operating in Azure or AWS, Rampiva will also spin down the machines that were running the finished job. This means there is little-to-no time that the machines are up (and incurring charges) but not actively working on some aspect of the Workflow.

3.5.3. Centralized Reporting

Rampiva Automate writes over 100 unique variables to a central database as jobs are completed, giving users access to a detailed and real-time reporting environment. Users may add steps to their Workflow that bring Nuix Case Item Data into this same repository or scan existing Cases for activity executed manually through Workstation. This comprehensive data set can be managed in an embedded Rampiva Automate database and accessed using OData or directed to a SQL database.



3.6. Agility

Accelerating set-up time and reducing the cost of deploying new environments can help clients scale to new regions with unique privacy regimes. Rampiva supports different scenarios that allow for expanding to additional regions, with various management and speed to deployment considerations.

3.6.1. Fully Isolated Template-Based Deployments

This scenario consists in performing an initial deployment of Rampiva Automate and storing this as a template, by creating templates of the virtual machines (VMs) used.

When a new region needs to be set up, the template VMs are instantiated in the new region, and minor configuration adjustments are performed post VM creation. A new fully configured environment can be set up in less than a day.

3.6.2. Fully Isolated Playbook-Based Deployments

This scenario consists of performing an initial deployment of Rampiva Automate and documenting the deployment steps in a playbook, as well as exporting Workflows and Profiles.

When a new region needs to be set up, the VMs and accompanying Azure or AWS resources are created in the new region, and the playbook is followed to complete the deployment. The new environment is then loaded with the previously exported Workflows and Profiles.

A new fully configured environment can be set up in less than a day.

3.6.3. Centrally Managed Deployment with Isolated Processing

This scenario consists of deploying a lightweight Rampiva Automate management instance in a Europe region, and defining dedicated Resource Pools within Rampiva Automate for each region. This scenario has the advantage that all configuration Profiles and Workflows are already in place when setting up a new region.

When a new region needs to be set up, a new Resource Pool with the appropriate configuration is created, and processing VMs are instantiated in the region from predefined templates. Additionally, a Data Gateway can be set up in each region to support scenarios where data uploaded by users never travels outside of the region. Although the Rampiva Automate management instance will hold metadata related to the projects, the actual content of files processed remains within the region where the Jobs are started.

A new fully configured environment can be set up in a couple of hours.





rampiva.com

© 2021 Rampiva Global LLC. All rights reserved.