CloudFabrix

E-BOOK | July 2021

Data Done Right for AIOps

With Robotic Data Automation (RDA)

Andy Thurai

Principal Analyst,
The Field CTO
thefieldcto.com

Success of AI/ML Projects Is Underpinned by Enterprises' Ability to Operationalize Machine Data at Scale

The shine on AI and ML projects has never been brighter. Smart companies are rolling out these technologies across various functional areas to create top-to-bottom growth, employee productivity, customer satisfaction, and to make IT operations more efficient. One such technology that is emerging as a strategic initiative is AIOps (artificial intelligence for IT operations); its goal is to make any IT environment efficient, smarter, scalable and resilient.

Simply put, AIOps is about using AI/ML to improve IT functions. The first step in any AI/ML project is getting data together. For AIOps, this means immense quantities of varied IT data – like events, service tickets, asset dependencies, metrics, logs, traces, event notifications, etc. Because of its complexity, only few companies are able to build proper "data pipelines" that can feed their AIOps system wholistically and on time, thus resulting in very accurate and timely results.

This is set to change with the new paradigm of **robotic data automation (RDA)**. RDA is designed to democratize the ability to build and manipulate data pipelines as needed using low-code/no-code bots, and to automatically integrate, collect and use the data. This will ensure faster value from AIOps implementations and extend the utility of it beyond the core AIOps use cases.

> **If AIOps is not given the right data, you will get garbage results. Fix the AIOps data pipeline first to get meaningful results.**
>
> **- Andy Thurai, The Field CTO**

In this eBook, we will discuss the following topics:

1. Why AI/ML projects fail
2. Key data challenges
3. What RDA is
4. How RDA helps address data problems
5. How RDA accelerates AIOps
6. Use cases and business value

The Problem

Modern IT systems are hard to develop and deploy and even harder to operate since their IT operations data gets very messy. There are multiple issues with data, from collection to processing to storage to getting proper insights at the right time. All this is addressed by AIOps, which has taken over the reins of several IT functions. However, AIOps just on its own can be another complex process at hand. Data for AIOps handled in a conventional way will increases the risk of long implementation cycles and ongoing maintenance of data quality and handling processes. Robotic Data Automation (RDA) fulfils the need for quality data with minimal maintenance cost.

This e-book highlights the importance of AIOps and the need to automate the DataOps and MLOps aspects of AIOps with a new paradigm called Robotic Data Automation (RDA) for achieving higher success rate and improved ROI

What is RDA?

Robotic data automation (RDA) is a new paradigm to help automate data integration and data preparation activities involved in dealing with machine data for analytics and AI/machine learning applications. RDA is not just a framework; it also includes a set of technologies and product capabilities that help implement the DataOps/MLOps automation.

Robotic data automation (RDA) helps organizations realize value from data faster by simplifying and automating repetitive data integration, preparation and transformation activities using low-code workflows and data bots, including AI/ML-bots.

What is AIOps?

AIOps is about using AI/ML to improve IT operations. Over the years, IT operations have gotten very complex with hybrid and multi-cloud operations managed by siloed teams. While the underlying use cases might vary depending on the enterprise, the common use cases include the following:

1. Root cause detection
2. Unified view of IT operations - Provide complete observability and insights for data collected across different clouds using different tool sets
3. Noise reduction
4. Intelligent alert monitoring and escalation
5. Predictive intelligence and outage prevention
6. Automatic remediation of IT issues
7 MTTR (mean time to resolution) reduction
8. Anomaly detection
9. Self-healing

There are other additional use cases that some consider fringe use cases. I wrote an article explaining most of them here.

Why do so many AI/ML projects fail?

I wrote a detailed analysis on this topic here. Based on my conversations with enterprise CxO level executives, and based on my number of years of experience in the field, almost every AI and ML project, AIOps projects included, struggle because of one issue: data. To be more specific, they struggle because of data pipeline problems. Almost every single large enterprise has a data problem.

Honestly, most of them have a problem because they have too much data, and they are not sure how and what to get out of that. In the data economy, data collection is not sufficient. It needs to be experimented by data scientists, and it should be properly used to deliver business insights to executives and business users. To get timely insights, the data pipeline - from instrumentation to data collection to data wrangling to data cleansing - needs to be solved by any AI/ML project to succeed. It consists of two components, one of that is solved by DataOps and the other by MLOps.

> "84% of the businesses have a major problem with non-optimized data warehouses, high storage costs of data because of too much data, outdated IT infrastructure, and manual or slow processes that do not meet business needs."

What is DataOps?

DataOps is the portion that helps fix data pipeline issues. It is the orchestration and automation of people, processes, and technology to quickly deliver trusted, unbiased, timely and high-quality data to AI/ML systems.

Data is an issue

Data is the new oil: it is dangerous if used in its raw form. Imagine the discovery, exploration, extraction, refining, and transportation process oil goes through before it reaches gas stations for consumption. Much the same way, data needs to be refined before it can be used. In addition, raw data also needs to be classified and labelled properly for AI/ML systems to understand it. AI doesn't know what to do with the raw data unless the AI systems are trained to recognized it properly. Data needs to be identified, classified, cleansed, and properly labelled before AI/ML systems can even touch it, let alone get insights from it. This mundane process of "data wrangling" is a big issue for a lot of enterprises. This is a big process that is costly, time consuming, error prone, and very slow. It needs to be properly automated to get the data pipeline feeding the systems for model creation and inference. I wrote an article in *Forbes* talking about those issues in detail here.

> "According to Forrester, up to 73% of company data is unused for analytics and insights.

Data collection

Given that every enterprise is moving everything to the cloud, the basic digital delivery architecture needs to be built for clouds from the ground up. A cloud-native, microservices-based, API-enabled, K8-based architecture is very complicated. Any mature digital native enterprise will have many hundreds of microservices that provide a cohesive digital application, and any mature digital native will make multiple changes per day per application. The total changes made can be in many thousands for a pure cloud native enterprise. Unfortunately, those changes are primarily responsible for unplanned outages the majority of the time. As the DevOps guru Gene Kim points out in his famous book *The Visible Ops Handbook*, 80% of unplanned outages are caused by changes. The changes can be either to infrastructure or to the application itself. To avoid this situation, the entire digital ecosystem needs to be instrumented for observability. This process needs to be automated, such that every container and every microservice that gets dropped in will have to come instrumented for data collection. The only way to properly integrate, automatically instrument, collect, and get the data to the right place will be by automation. A manual process of data collection will have data blind spots that will result in partial observability that produces skewed insights.

Data integration

When the data collection is siloed, and when storage is siloed, it becomes difficult to integrate the data at source or with data from multiple sources. Data integration is one of the most time-consuming, expensive, and error-prone tasks. The majority of organizations bring professional expertise, such as SIs or consulting companies, to integrate data producers into your systems. Initial integration is expensive and time consuming, but if the integration is done through code, maintenance of that code becomes very difficult. Any time a system on either side changes, there will be a problem. One way to eliminate this is to API-enable the data producers and data consumers. This allows for somewhat of an easier integration, but it still is an involved and time-consuming task, especially since the input is a combination of structured, unstructured, and semi-structured messy data.

Data preparation

Data preparation is NOT an easy task. To achieve the best possible results in the data economy the data needs to be in its best possible form. The “dirtier” the data, the more skewed the results will be. Dirty data can have duplicate data, missing information, inaccurate information, incorrect information, or even biased information.

> “ On average, about 80% of a data scientist’s job is spent on preparing the data.

On average, about 80% of a data scientist’s job is spent on preparing data. Raw data in its inherent nature is very dirty, and at least 20% of the raw data is said to be “dirty.” The accuracy of the AI/ML model depends on the accuracy of the data. If the data is not prepared properly, it will skew the model. If the model is wrong, then the insights, decisions and recommendations will all be wrong. For the most part, AI is based on data. I wrote a piece on that in *Forbes* about a year ago titled ”Why Infusing AI into IT Operations Is More about the Data Than about AI Itself” where I discuss this in detail. If data is not prepared properly, it will result in a “garbage in, garbage out” situation.

This is an area where RDA can help a lot. Automatically have the robots clean the data to get it to its usable form.

After getting the raw data, it needs to be contextually enriched with related/adjacent information as well. For example, what else happened at that time? Were there any anomalies or any events that triggered it? Was the system over capacity? Did the feeder systems go down?

Data bias

The industry is starting to deal with this issue --avoiding bias while making decisions or drawing insights. There are about [180 human biases]() that have been identified and classified, and human processes and procedures are in place to mitigate them. Unfortunately, many of them are making their way into AI design and raw data, and the IT industry is not mature or ready to deal with them yet. Strangely enough, companies are willing to trust mathematical models because they “assume” AI will eliminate human biases; however, these AI models can also introduce a set of biases if they go unchecked. The #1 area that needs to be addressed is the automation of the data cleaning to eliminate bias before the models are created. Eliminating bias from the datasets is an easy task that RDA can easily handle. If the AI model/algorithm is based on biased data, then the decision will obviously be biased. The only way to eliminate this problem would be to analyse the input data for inequalities, biases, and other negative information which can lead to fair decisions. The models need to be updated on a constant basis, and the data that feeds those models needs to be cleansed as it comes to it. The only way to achieve this would be by automating the data cleansing process using RDA. [This article]() discusses more about human biases and how to eliminate them from AI models.

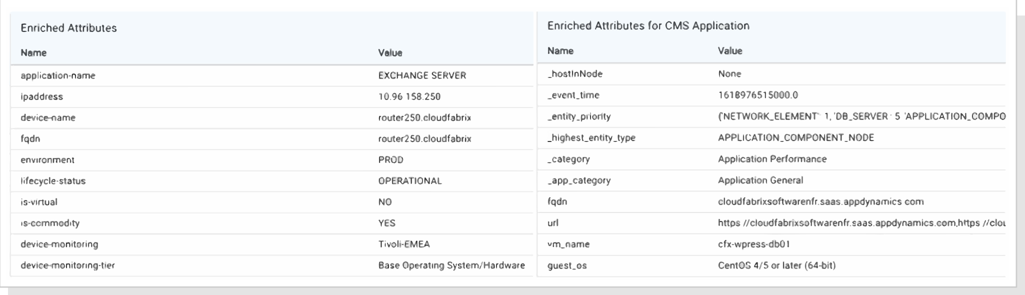
> “Most organizations spend a lot of time on data prep but mainly concentrate on preparing the data format and quality for consumption, but not on eliminating bias data.

Data issues in general

In general, AI/ML data-related issues can be classified into 3 groupings: getting the data from the source where it is produced, cleansing the data before it can be used, and enriching the data. After data is collected, integrated, and refined, it still is not ready to be used. This is where the labelling, classification, grouping, etc. comes into place. Data also needs to be correlated. Data enrichment in general is about taking the raw, cryptic data and adding information to it so it can be helpful to support teams in case of support issues or data scientists and/or decision makers to make the right decisions. Bots can do data enrichment easily; the manual process to do the same can be expensive.

> “2,700+ LOB executives report their top AI initiative as “modernizing data infrastructure for AI” as per State of AI in the Enterprise survey from Deloitte Insights.

Data Enrichment

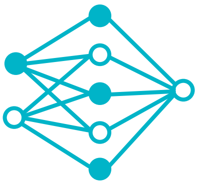


Enriched Attributes

Name	Value
application-name	EXCHANGE SERVER
ipaddress	10.96 158.250
device-name	router250.cloudfabrix
fqdn	router250.cloudfabrix
environment	PROD
lifecycle-status	OPERATIONAL
is-virtual	NO
ıs-commodıty	YES
device-monitoring	Tivoli-EMEA
device-monitoring-tier	Base Operating System/Hardware

Enriched Attributes for CMS Application

Name	Value
_hostInNode	None
_event_time	1618976515000.0
_entity_priority	('NETWORK_ELEMENT' 1, 'DB_SERVER · 5 'APPLICATION_COMPO
_highest_entity_type	APPLICATION_COMPONENT_NODE
_category	Application Performance
_app_category	Application General
fqdn	cloudfabrixsoftwarenfr.saas.appdynamics com
url	https //cloudfabrixsoftwarenfr.saas.appdynamics.com,https //clou
vm_name	cfx-wpress-db01
guest_os	CentOS 4/5 or later (64-bit)



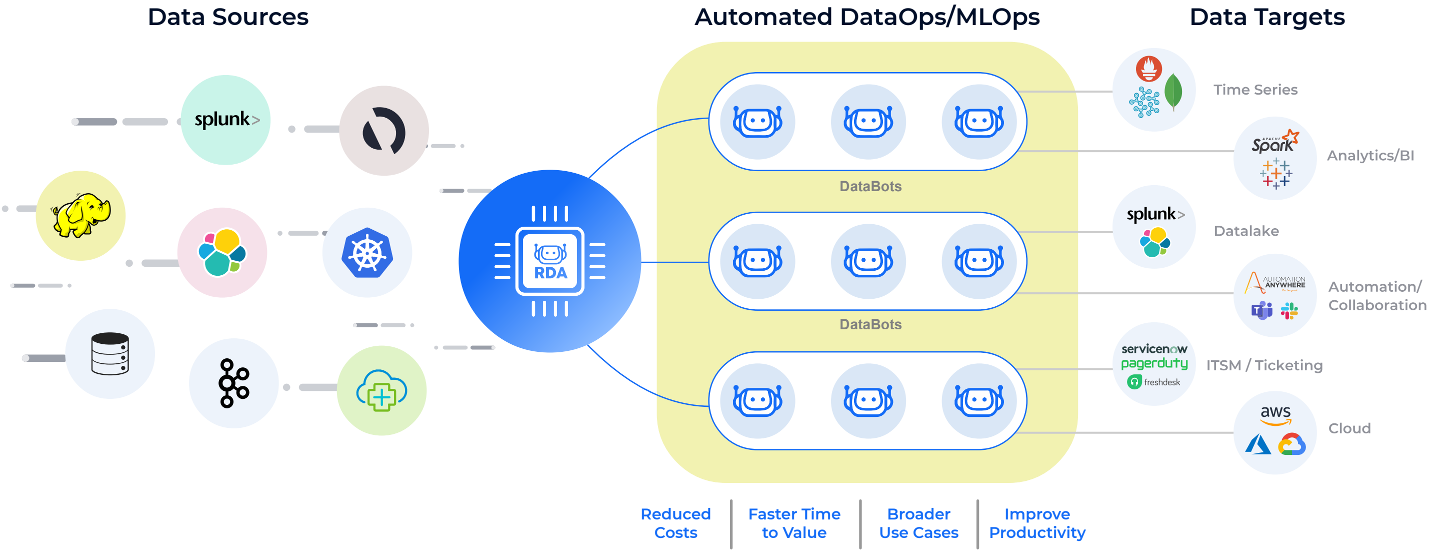
> RDA is designed to automate the DataOps with a framework and an array of software bots

Achieve DataOps automation

As is evident from the above, data operations are still a big problem with enterprises. missing information, inconsistent information, errors, omissions, or even bias information can cause datasets to not be useful for AI/ML projects. A clean, holistic and automated DataOps is required at the foundation of all AI/ML projects.

RDA is designed to automate the DataOps with a framework and an array of software bots that help accelerate and simplify all data handling required for AI/ML projects. It provides an automated and consistent approach for data collection, integrity checks, and data cleansing, transforming and shaping the data (aggregating, filtering, and sorting) for consumption by the AI/ML systems.

We will discuss this new paradigm in detail in the subsequent sections.



Principles of RDA

Robotic Data Automation (RDA) is a set of robotic processes that are used to solve some of the above data pipeline issues. Robotic Process Automation (RPA), on the other hand, is used to solve process and workflow issues. RDA is generally a DataOps automation framework with bot libraries and data pipeline management tools. Generally, it is about automating the manual, cumbersome, expensive data pipeline process with bots.

Key RDA principles that set it apart from ETL approaches

01

Low-code/no-code approach

It is important to be able to build data pipelines with a low-code/no-code approach. Customers should be able to build data pipelines using IDE style interface. A key aspect with "Low-Code/No-Code" pipelines is that you don't have to know Python or any scripting language to implement these pipelines. Pipelines in RDA just use configurational semantics using text that is similar to natural language.

02

Universal query language: Uniform interface for every bot

All integration and data bots must offer a uniform interface to perform DataOps activities like data collection, performing integrity checks, data cleaning, transforming and shaping the data (aggregating/filtering/sorting). This consistent interface makes it easy to handle any type of data from any end point, minimizing the specialized domain skills needed for the pipeline developers.

03

Apply DevOps model to DataOps pipelines

RDA must provide DevOps-like agility for data pipeline developers, enabling rapid prototyping, iterative experimentation, and collaboration. RDA should provide a collaborative visual workbench or IDE, using pipelines that can be built or customized and visualized; once finalized, they can be pushed on to the production AI/ML system, e.g., an AIOps platform. Enterprises can easily integrate with existing CI/CD processes to verify and publish the pipelines.

04

Pipeline tracing, data lineage and accounting

You need to have built-in capability to track and log every change to the data within a pipeline and across pipelines. Also, you need the ability to check the input and output at every step of the pipeline. This is very useful not only for debugging but also for tracing the lineage for compliance and accounting purpose. It should provide a visual interface to easily understand the data flow and pipeline execution sequence.

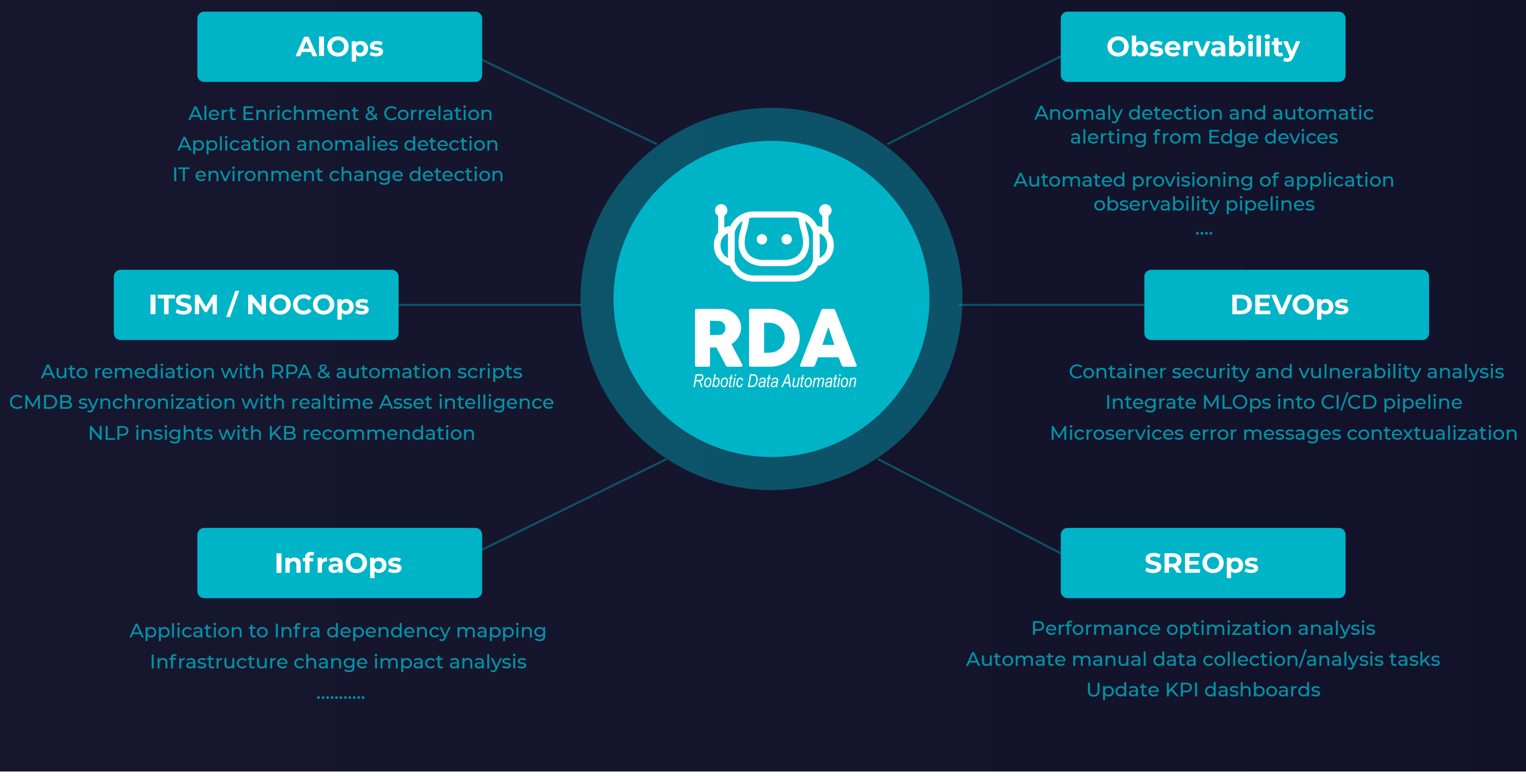
05

Data governance: Control over what data can go where and retention

Data governance is an essential part of the DataOps functions. It helps the organization know what data you have, where that data resides and how that data is being used. RDA should provide roles-based controls to authorize who can access what data at what time. It should also provide a visual interface to check how the data is moving between various end points. The detailed logging of data operations can be used to verify any policy deviation.

Broad Set of Use Cases Beyond AIOps

RDA lets you get to AIOps insights faster, and it makes AIOps more open and extensible and allows you to work with data more efficiently. As a result, you can solve many IT use cases across multiple IT disciplines such as Observability, ITOps, ITSM, DevOps and more.



AIOps

Alert Enrichment & Correlation
Application anomalies detection
IT environment change detection

Observability

Anomaly detection and automatic alerting from Edge devices

Automated provisioning of application observability pipelines
....

ITSM / NOCOps

Auto remediation with RPA & automation scripts
CMDB synchronization with realtime Asset intelligence
NLP insights with KB recommendation

RDA
Robotic Data Automation

DEVOps

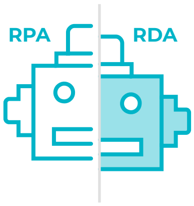
Container security and vulnerability analysis
Integrate MLOps into CI/CD pipeline
Microservices error messages contextualization

InfraOps

Application to Infra dependency mapping
Infrastructure change impact analysis
....

SREOps

Performance optimization analysis
Automate manual data collection/analysis tasks
Update KPI dashboards



RDA and RPA: Better together

How does RDA compare and differ from RPA? RPA automates business processes and user tasks, whereas RDA automates data tasks. RPA provides software bots to mimic user actions. However, this has not eliminated or reduced the tasks related to data handling from various tools and systems that IT is responsible for. Moreover, these tasks continue to get harder because modern complex systems are a lot more dynamic with a distributed nature to produce vast amounts of data at a much faster rate.

This is the problem that RDA is designed to address. RDA is both a data automation framework and a toolkit to accelerate and simplify all data handling of IT. RDA is generally DataOps that streamline and automate an associated MLOps framework with bot libraries and data pipeline management tools. Generally, it is about automating the manual, cumbersome, expensive data pipeline process with bots.

In summary, RDA automates DataOps and associated MLOps, similar to what RPA did to automate business processes. RDA perfectly complements RPA by automating complex data workflows and integrations.

> “ Simply put, RPA automates business processes and user tasks, whereas RDA automates data tasks.

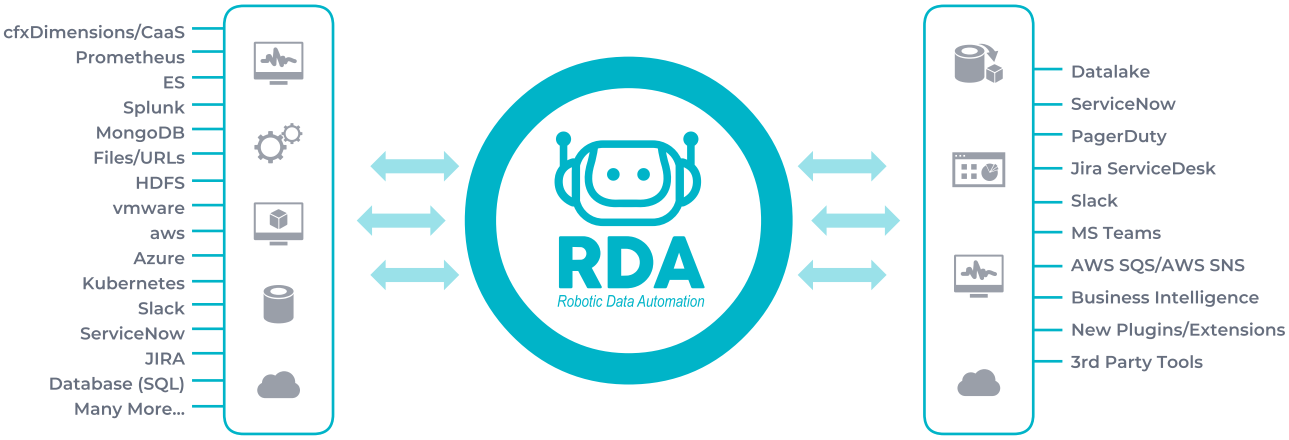
RDA Business Value to Enterprises

RDA can help businesses jumpstart and accelerate their AI/ML journey. A few key benefits enterprises can realize with RDA enabled AI/ML projects are outlined below:

1. **Faster ROI:** Realize AI/ML benefits faster
2. **Scale exponentially:** When enterprise data grows exponentially, it is easy to scale up RDA with bots versus manual or semi-automated scale up The insights, decision making, and overall business can grow hyper scale with the data volume.
3. **Cost reduction/containment:** Reduction in cost and man hours spent due to less reliance on manual activities
4. **Eliminate costly errors and biases:** Enterprises can identify and eliminate errors, biases, and other quality issues much before data scientists get to work.
5. **Faster project proposals:** Enabled by exploratory data analysis, what-if analysis, pre-trained bots, etc.
6. **Improved IT productivity:** Field teams, implementation, services, and support teams will be more productive.

Operationalizing the machine data at scale is not an easy task. Different sources, different types of data, varying volumes, and varying velocity can pose interesting challenges. Automating the data collection, data integration, and data preparation using AI/ML bots can power the AI systems to learn on the job either supervised or unsupervised, adjust models on the fly, and work more efficiently. Obviously, this helps enterprises realize value from data faster by simplifying and automating repetitive data ingestion from various sources, data preparation, and data transformation activities with low-code or no-code pipelines and recipes built with AI/ML bots.

Data tasks that can be automated using RDA include data collection, data integration, data validation, data clean-up, data normalization, meta data enrichment, and data extraction from structured or unstructured data. Getting the value out of data is still a major issue for non-digital native enterprises. Most organizations think having too much data means they can get a lot of value from that data. Unfortunately, it doesn't work like that. Volume never equates to value when it comes to data. It is the quality of data that matters. If these mundane, time-consuming, and costly tasks can be automated, then any enterprise can realize ROI a lot quicker and TTM will be much faster.



Conclusion

In the AI/ML data pipeline, automating everything is possible. To make it more efficient, AI needs to be used to help the AI systems. Automate everything that is mundane, is repeatable, and produces measurable results. DevOps brought a new coding culture and discipline to the compute/code-based industry. Use DataOps to bring data discipline to the data economy.

All the so called “digital-native” companies have figured this out because they all thrive on data. The successful ones have figured out how to handle the data pipeline properly and how to scale up with data volume uptick.

The bottom line: if you want to become a “data-driven” company, fix your data issues first.

Additional Resources

RDA Webinar with Andy Thurai, Principal, The Field CTO.
https://www.youtube.com/watch?v=Jy6ablWbxWQ

AIOps has a Data(Ops) problem
https://thefieldcto.com/aiops-has-a-dataops-problem/

Robotic Data Automation (RDA)
https://www.roboticdata.ai/

RDA Bot Library
https://roboticdata.ai/bot-library/

RDA AIOps Studio | Free Signup & Getting Started
https://www.youtube.com/watch?v=Js67MUfNZHM

www.cloudfabrix.com
info@cloudfabrix.com