RADIANCE

A COMPREHENSIVE INTRODUCTION

How Lumina Has Optimized Its Artificial Intelligence and Proprietary Technology to **Protect What Matters Most**

FEBRUARY 2021

This document and information herein contains confidential, proprietary information and is subject to the Non-Disclosure Agreement (“NDA”) entered into between Lumina Analytics, LLC and the recipient and may not be disclosed to third parties or published pursuant to the terms of the NDA.

Lumina

TABLE OF CONTENTS

CONFIDENTIAL AND PROPRIETARY

INTRODUCTION

Imagine a system that can take any size population of individuals, conduct a full internet search, and rank the individuals by percentile in order of probable behavioral affinity. That is what we have created with Radiance.

Radiance is a due diligence solution designed to aid human analysts in identifying threatening behavior.

The platform operates at unmatched speed and scale in a highly parallelized system of unique, proprietary, and interdependent self-learning algorithms. These algorithms include data mining, data integration, natural language processing, name extraction, identity resolution, and behavioral models, all of which are used to generate a summarized and prioritized report for the given population.

The purpose of this paper is to provide a technical overview of Radiance's computational pipeline and to summarize the value proposition associated with each element of its intellectual property.

The purpose of this paper is to **provide a technical overview of Radiance's computational pipeline** and to summarize the value proposition associated with each element of its intellectual property.

The Problem

As part of necessary due diligence processes, organizations and governments commonly utilize the open internet to identify and minimize risks. However, the massive growth in online data and the restrictions posed by traditional search engines create challenges. Analysts commonly conduct due diligence on large populations by manually searching for risks on traditional search engines. This approach to due diligence is costly, incomplete, prone to human error, and susceptible to bias introduced by search engines themselves, which cater to users as consumers.

The 5.32 billion pages of information that exist on the internet today is expected to continue growing exponentially as our world becomes more interconnected, digitized, and driven by artificial intelligence.1 Without a more sophisticated due diligence solution, the quantity of data necessary for thorough due diligence exceeds the capacity for human analysis. In addition, a manual approach makes it impractical to monitor populations over time and at scale.

The **5.32 billion pages of information that exist on the internet today** is expected to continue growing exponentially as our world becomes more interconnected.

[1] "WorldWideWebSize.Com | The Size of the World Wide Web (The Internet)," accessed January 14, 2021, https://www.worldwidewebsize.com/

RADIANCE AS THE SOLUTION

One search in Radiance is equivalent to more than 50,000 queries using a traditional search engine

Radiance is a software-as-a-service (SaaS) platform that automates searches across the internet and proprietary data assets, to identify risk quickly and precisely at scale. To our knowledge, as of January 2021, Radiance is the only platform in existence that can ingest the entire set of relevant pages about an individual/entity on the indexed internet and provide unbiased search in minutes. In fact, one search in Radiance is equivalent to more than 50,000 queries using a traditional search engine.[2] Therefore, by using Radiance, a single analyst can complete in minutes what would have taken years using the traditional search methods.

Radiance does this by starting with a name or list of names of individuals. The platform allows analysts to add any identifying data points related to each individual to assist with identity resolution. It then collects all data available on the open internet and from external data sets related to individuals within a given population. It steps through a series of algorithms to verify the page presence of each individual and scans the data sets for threatening behavior. If a population is monitored over time, Radiance has the capacity to analyze only what is new within each data source.[3] Following these processes, the platform organizes its findings by ranking individuals according to the behaviors found, and as a final step, the system generates a completed report that neatly summarizes those findings. The report may be edited within the platform or exported as is.

Radiance achieves its speed because of how it approaches the problem.

[2] For details, please see note 7.

[3] Lumina calls this a “delta search.” Radiance provides the option of a repetitive search that monitors the population of individuals with a predetermined frequency. Delta searches show changes in behavior over time, indicating new potential risks appearing since the last delta search date.

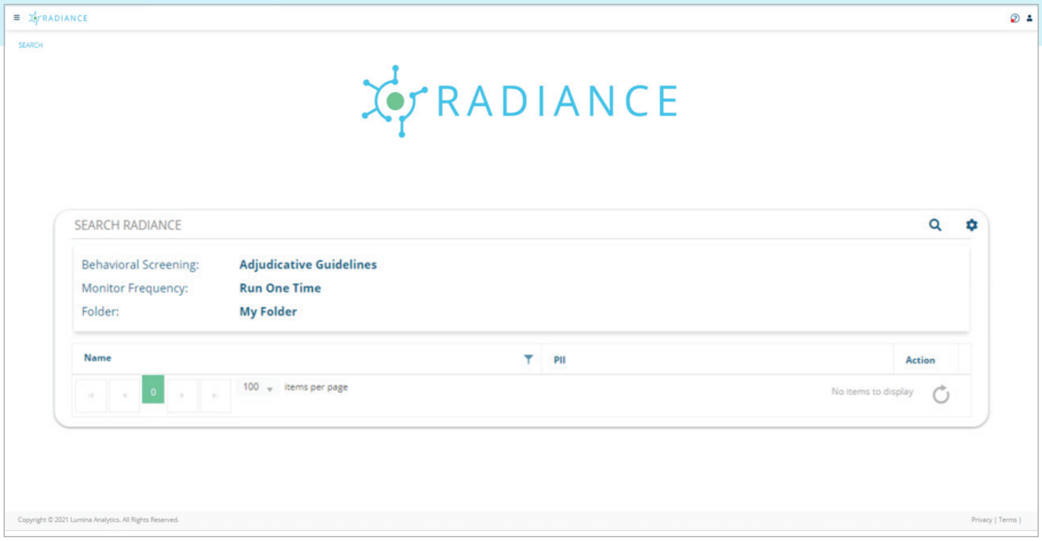
The Computational Stages of Radiance

Radiance search begins by entering a name (see Figure 1).

FIGURE 1 I

Initiating a Radiance Search

Figure 1 displays a user typing a name into the search bar, selecting a frequency to monitor the name, and selecting the behaviors to screen.



Behavioral Screening: Adjudicative Guidelines
Monitor Frequency: Run One Time
Folder: My Folder

Clicking search begins the following sequence:

1. URLs that are indexed by the given name are collected.
2. The resulting list of URLs is distributed to a large set of servers.
3. The servers scrape each URL and stream the downloaded pages to a central cluster for behavioral analysis.
4. Pages without relevant phrases are discarded.
5. The pages that remain are linked with their corresponding names and identifying information via identity resolution.
6. These pages are then ranked following a predetermined model that is produced by machine learning.
7. Each page is finally summarized, deduplicated, and presented in an automated report.

Addressing Biases

The internet is commonly misconceived as a single, massive collection of documents. In contrast, the internet is better understood as a collection of algorithms that seek to display popular and profitable content.[4] Because of this dynamic, information essential to due diligence processes can become buried in the deep web. In addition to the biases created by traditional search algorithms, human analysis introduces confirmation bias. Confirmation bias occurs when an analyst identifies patterns that are expected or desired but not really present.

By analyzing massive data sets exhaustively, Radiance eliminates the potential for confirmation bias and search engine bias.

By analyzing massive data sets exhaustively, **Radiance eliminates the potential for confirmation bias and search engine bias.**

Integrating Other Data Sources

Since its founding, Lumina has collected and procured several proprietary datasets representative of various threatening behaviors and has the capability to ingest third party data. Any of these datasets can be added to a user's Radiance configuration. Regardless of how many datasets are added, all data is processed in parallel and presented in one automated report once the sequence of algorithms is complete.

Lumina has developed a high-speed proprietary in-memory indexing method for its internal databases. When comparing Lumina's in-memory solution to similar in-memory database products, Lumina's database technology performs 10 to 50 times faster when employing algorithmic settings for typical and maximal parallelization (see Figure 2). This level of performance allows Radiance to maintain its speed regardless of the size of the dataset processed. In addition to outperforming Arango DB, Radiance displays close to linear behavior as the complexity of its queries grows, the implications of which lie beyond the scope of this paper.

[4] The PageRank algorithm used by Google today prioritizes pages that are linked most often by other pages.

[5] An N-gram is a sequence of N words. E.g., "one hundred queries" is a trigram.

[6] Many websites have security to prevent DoS attacks. Radiance avoids triggering these cyber defenses while maintaining its speed and efficiency in gathering data.

FIGURE 2 |

Lumina's Data Lake Indexer (multiple algorithmic settings) vs Arango DB

The test behind **Figure 2** employed Lumina's proprietary dataset of documents related to international terrorism. Figure 2 compares Lumina's Data Lake Indexer, at different settings for number of queries executed in parallel, to Arango DB with respect to how quickly each identified all documents in the dataset containing each of 100 given n-grams.[5]

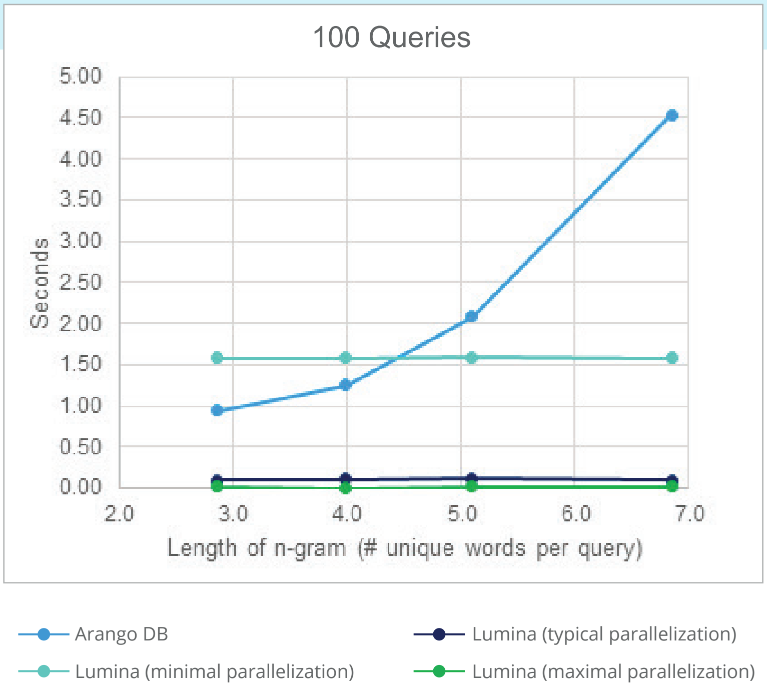
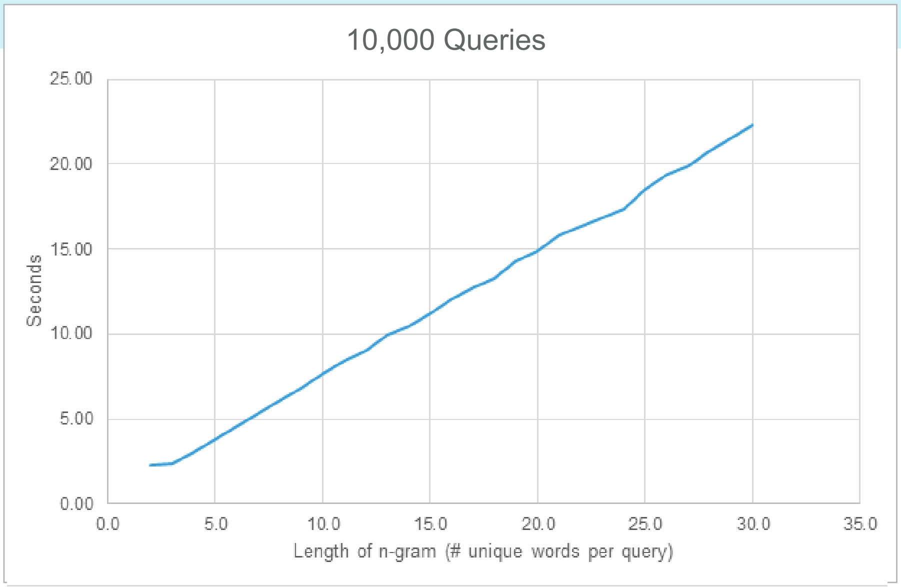


FIGURE 3 |

Stress Test of Lumina's Indexing Algorithm (maximal parallelization)

Figure 3 illustrates Lumina's proprietary database performance as complexity of queries increases dramatically. The time required increases linearly.



Ingesting the Internet at Scale

Lumina's internet ingestion orchestration engine treats the internet as a distributed database. It has been developed both to ingest data as quickly as the internet will allow and to maximize the throughput of Radiance's analysis cluster. Should a server fail, the engine is self-correcting. It recovers any computational loss and brings the server back online. These safety and stability measures may be monitored and controlled from a central command module. As each server collects data, it streams its data to the analysis cluster and then listens for its next instruction.[6]

Behavioral Affinity Models

A Behavioral Affinity Model (BAM) is a set of phrases that indicates a specific behavior. As pages of data pass through the analysis cluster, pages that include BAM phrases are retained for further processing and pages without BAM phrases are discarded.

The set of phrases that makes up a BAM may be produced by Lumina's BAM Generator algorithm or by subject-matter experts. In both cases, the lists of phrases are readable by analysts and may be adjusted at any time. Alterations to a BAM can be communicated to the analysis cluster without downtime and will take effect for all Radiance queries following the change.

Identifying risk and relevance through Behavioral Affinity Models lies at the heart of what makes Radiance powerful and unique.[7]

Identity Resolution

Radiance employs a method of identity resolution that allows it to detect variations of names. These include variations in the order of and distance between words that make up a name, e.g. whether the first or last name appears first or whether an affix interrupts the continuity of a name. In every case, Radiance links the name to its accompanying data and the specific behavior searched. If the name of an individual is unknown or ambiguous due to variations in translation or transliteration, Radiance treats each variant as an individual name within the population. In its final stage of identity resolution, Radiance uses identifying data points to further filter pages that display the name and specific behavior searched.

Generating Models with Machine Learning

Lumina's proprietary BAM generator utilizes a genetic algorithm to identify the minimum number of phrases (also called n-grams) needed to distinguish two collections of text. In other words, the algorithm generates the best and most efficient model for isolating the behavior reflected in the set of documents. The algorithm uses a proprietary in-memory optimization that allows for it to learn quickly and to produce candidate phrases efficiently.

Lumina produces new behavioral models with two methods:

1. contrasting datasets reflecting a specific behavior (training datasets) with data reflecting no specific behavior (neutral datasets), and
2. contrasting datasets reflecting a specific behavior (training datasets) contrasted with datasets reflecting a different specific behavior (alternate training datasets).

When Lumina created its behavioral model for suicide, it mimicked an experiment conducted by the University of New South Wales.[8] Researchers at the university achieved an accuracy of 76% when constructing a corpus to distinguish text indicative of suicidal behavior from neutral text. Lumina's BAM generator achieved 81% accuracy by contrasting Lumina's proprietary training dataset with a neutral dataset.

In many respects, Radiance is agnostic to language. It is currently configured and has been thoroughly tested to run with BAMs containing phrases in English and to score risk in accordance with these. However, experimental evaluation in Spanish, French, Portuguese, and German suggests that Radiance can generate BAMs and screen names in any spelled language.

[7] As of January 2021, Radiance screens for just over 3,200 BAM phrases. Radiance also searches for each name in combination with 16 other key terms. Therefore, a single search in Radiance is equivalent to more than 3,200 x 16 = 51,200 searches using a traditional search engine. In addition, Radiance analyzes far more links per search than a human analyst could.

[8] McHugh CM, Large MM. Can machine-learning methods really help predict suicide? Curr Opin Psychiatry. 2020 Jul;33(4):369-374. doi: 10.1097/YCO.0000000000000609. PMID: 32250986.

Percentile Scoring

Radiance screens names of individuals against Behavioral Affinity Models. Each page that Radiance collects is analyzed for the name of the individual, identifying data points of the individual, and phrases that indicate a specific behavior. As relevant pages are identified for an individual, Radiance counts the number of pages found for each behavioral model. The final number of pages found for each model creates a signature for the individual. This signature is fed to the Percentile Scoring algorithm which compares it to the signatures of all other individuals in the given population. It then calculates a risk value that is used to sort the highest-risk signatures to the top of the list of names so that they may be presented first in the automated report. As of January 2021, the Percentile Scoring algorithm is capable of ranking individuals with 88% accuracy.

FIGURE 4 |

The Evolution of Lumina's Percentile Scoring Algorithm

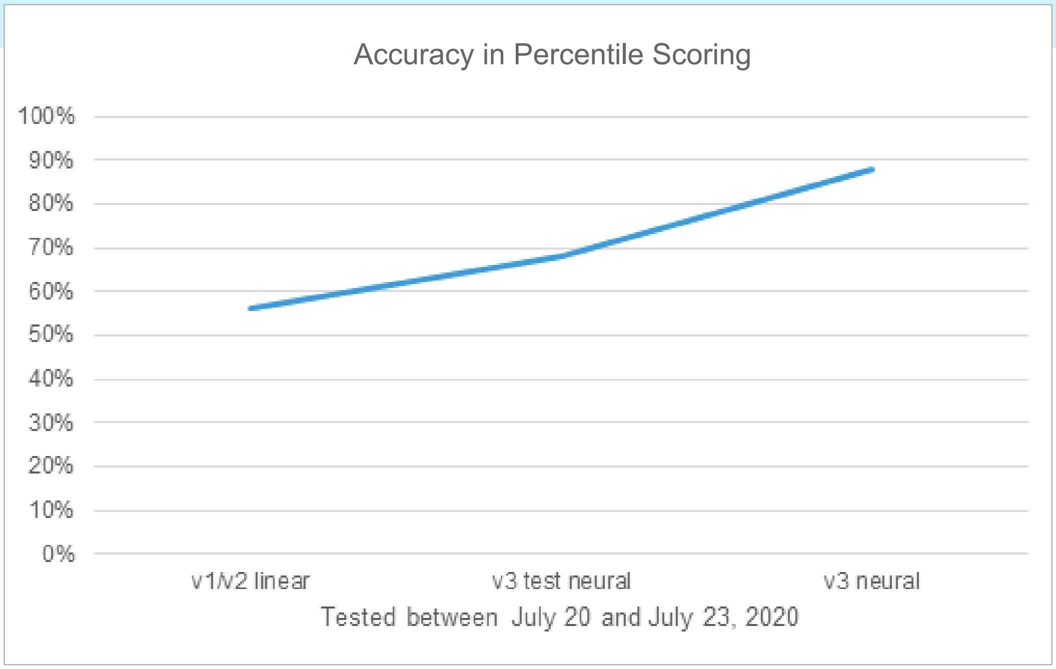


Figure 4 illustrates the iterative improvement of Lumina's Percentile Scoring algorithm – measured as the ability to rank names within three percentiles of the rank assigned to them by a human analyst. The test behind Figure 4 used the algorithm's own training data and consequently demonstrates the accuracy of its recall. The algorithm utilizes signatures from all Radiance searches to learn continuously and to improve over time.

Name Extraction

If identifying risk through Behavioral Affinity Models lies at the heart of Radiance, the Name Extraction algorithm (NxSM) lies at the heart of Lumina. The algorithm inspired much of the development of Radiance. The algorithm is foundational for Radiance's Auto-Summarization algorithm and Deduplication algorithm, both discussed below.

Lumina's Name Extraction algorithm identifies the names of individuals from text documents and has been designed from the ground up efficiently to process many formats including HTML, JSON, and CSV files. The algorithm outperforms popular entity recognition tools when tested on the CoNLL-2003 standardized entity recognition benchmarking test.[9] Additionally, Lumina has tested its Name Extraction algorithm on case-less documents to illustrate the algorithm's stability (see "Name Extraction Accuracy" below).

TABLE 1 | **Name Extraction Accuracy**

	Clean	Lowercased
Lumina	**95.3%**	**93.1%**
Stanford NER	**89.8%**	**0.0%**
SpaCy	**76.9%**	**0.0%**
NLTK	**72.6%**	**0.4%**

Table 1 demonstrates the stability of Lumina's Name Extraction algorithm. When identifying the names of individuals, Lumina's algorithm outperformed StanfordNER, SpaCy, and NLTK on a collection of predictably-cased text documents. The text was then lowercased to mimic social media and other informal online content. Lumina's Name Extraction algorithm maintained its accuracy, whereas the competing tools became completely ineffective. In addition to achieving better accuracy than competing entity recognition tools, Lumina's algorithm processes data 300 to 350 times faster than modern comparable algorithms.

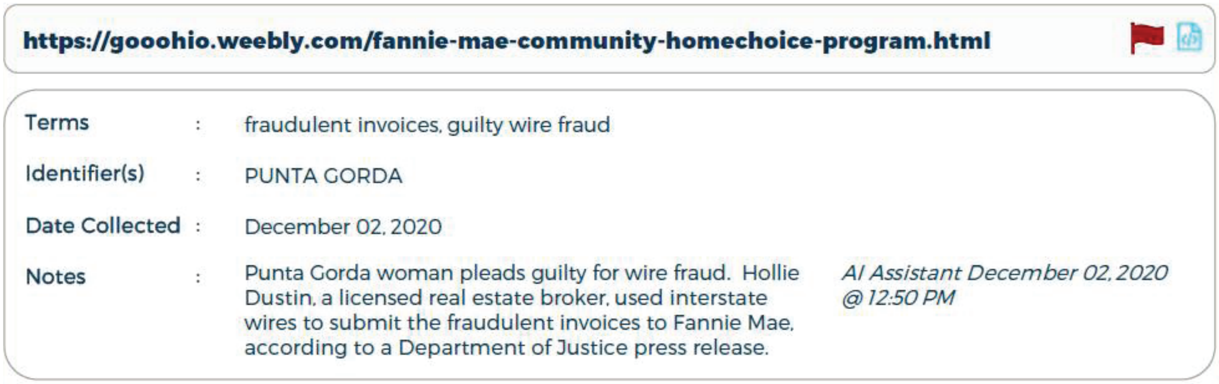
[9] Cf. https://www.clips.uantwerpen.be/conll2003/ner/

Auto-Summarization

The Auto-Summarization algorithm provides succinct, actionable information for Radiance's automated reports. The algorithm does so by collecting the text surrounding the name screened and the specific behavior identified. These lines are saved and included in the automated report (see Figure 5).

FIGURE 5 |

Example of Auto-Summarization



https://gooohio.weebly.com/fannie-mae-community-homechoice-program.html

Terms : fraudulent invoices, guilty wire fraud

Identifier(s) : PUNTA GORDA

Date Collected : December 02, 2020

Notes : Punta Gorda woman pleads guilty for wire fraud. Hollie Dustin, a licensed real estate broker, used interstate wires to submit the fraudulent invoices to Fannie Mae, according to a Department of Justice press release.

AI Assistant December 02, 2020 @ 12:50 PM

Deduplication

Radiance also uses the Auto-Summarization algorithm to deduplicate redundant content. If multiple URLs produce pages with identical summaries, duplicates are discarded and only one of the URLs is presented in the report. Duplication commonly appears when scraping websites that curate news from the same sources.

Deduplication ensures that individuals are ranked appropriately and reduces the workload of an analyst consuming the automated report. As with all algorithms mentioned, the Deduplication algorithm occurs in-memory and is optimized for its purpose within Radiance.

Lumina's Alternative Back-Propagation Method

Each of the algorithms mentioned in this section benefits to some degree from Lumina's Deep Learning Neural Network. In addition to developing each model, Lumina has created an alternative approach to back-propagation, nicknamed "Edison," that effectively trains neural networks nearly twice as fast as traditional methods.

Performance

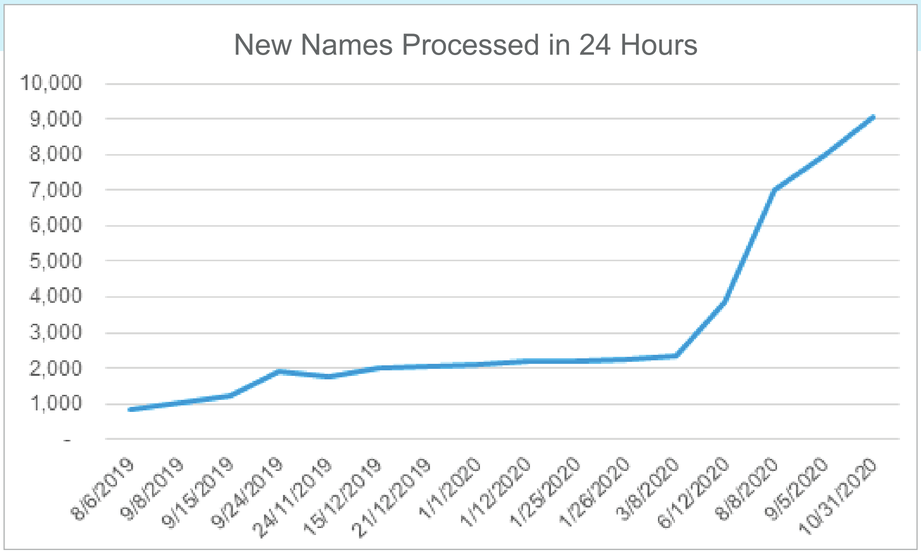
As of January 2021, each instance of Radiance consists of 330 physical machines that execute the algorithms explained above. In this configuration, each instance is capable of consistently processing 8,000 new names per 24-hour period with speed peaking around 9,000 names.[10]

FIGURE 6 |

Radiance Processing Performance

Figure 6 illustrates the increase in speed that Lumina achieved over a year of development.

Lumina's approach to improving speed has been methodically to identify and to address bottlenecks throughout the complex system.



Each performance test followed a set of code changes and replacement of legacy code. Stability also improved dramatically in June 2020 after the transition from MSFT Azure to the IBM Cloud.

Efficacy

In addition to tracking performance metrics, Lumina continuously monitors Radiance's quality of output. A team of analysts label output data as indicative of risk or not indicative of risk and record their findings. This dataset acts as a control group for monitoring Radiance output, which ensures that neither modifying Behavioral Affinity Models nor tuning of the algorithms increases false negatives.

Scalability

Radiance is linearly scalable. In other words, as demand for search increases, Radiance can grow, in its current configuration, in increments of 330 machines, or 8,000 names per 24-hour period. As Radiance grows, operations remain simple because the collection and analysis of data may be distributed to any instance of the system and produce equivalent results.[11]

[10] "New" names are names that Radiance has not searched previously. If a name is submitted that is identical to a name searched previously, Radiance analyzes old and new data in parallel. Because less data is processed each time the same name is searched, the speed with which a population is searched iteratively improves over time.

[11] In determining the number of instances required for a specific task, any fraction of an instance must be rounded up to the next whole instance, in order to preserve speed and efficiency optimization.

CONCLUSION

Lumina has developed all of its algorithms and intellectual property specifically for Radiance as a due diligence solution. This has led to the creation of an extremely precise system that is optimized well beyond commonly used and generic, open-source algorithms.

In addition to their speed, Lumina's algorithms improve over time as a product of the data they collect. The circular training of the data improving the algorithms and the algorithms improving the quality of the data results in a system that is both unique and irreplicable.

Data has been collected and algorithms have been tuned continuously since Lumina was founded in 2015. Because Radiance exists solely as a cloud offering, no organization has access to all of the individual pieces of the platform. The system supporting the Radiance platform is sufficiently vast to require 24-hour cycles for tuning. These factors render the platform difficult for potential competitors to imitate.

Lumina's Radiance introduces a new way to identify threatening behaviors within large populations. As data continues to grow rapidly, Radiance will address the world's need for a modern due diligence solution.

As data continues to grow rapidly, **Radiance will address the world's need for a modern due diligence solution.**

RADIANCE

Lumina

Lumina Analytics, LLC
501 E. Kennedy Blvd. Suite 810
Tampa, FL 33602

Email us at Support@Lumina247.com