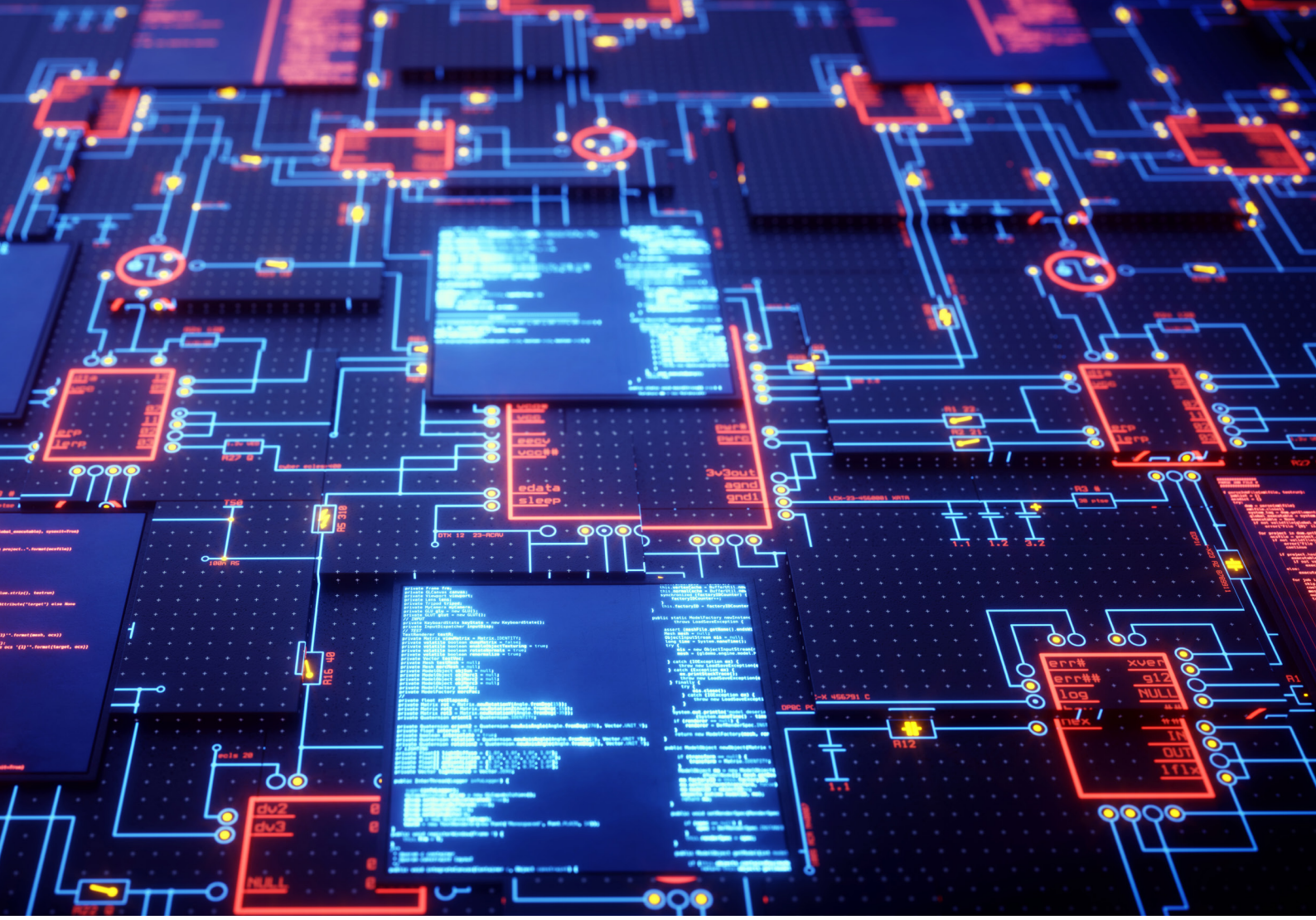

WHITE PAPER

Escalating Risk in Document Archives for GRC

BY

BORIS KHAZIN

Director, Governance, Risk & Compliance, EPAM

JAMES BUONOCORE

Business Solutions Consultant, EPAM

Table of Contents

1. | When Risk Emerges from Within

Risk has a habit of culturing itself in the most unlikely places. Worse, change can introduce risk to previously innocuous parts of your business. One example includes the intensifying concerns around data privacy and large security breaches, which have fostered new legislation, regulation and litigation, often with significant monetary and reputational damages. Businesses have taken the baton, working hard to seal off every vulnerability from external threats, but what if the risk is already there, hiding in places you'd least expect?

Consider the enterprise document archive. Paginated documents are the life blood of governance, risk and compliance (GRC). Even in this digital-first age, they are still a basic necessity. This paper discusses how contemporary content processing solutions can transform information on physical-turned-digital pages into insights for companies and mitigate data risks for GRC. Specifically, we will look at personal identifiable information (PII) in archival documents—an often unnoticed and growing snowball of risk. At the end, we'll offer a short menu of additional use cases that would benefit from a comprehensive content processing solution. The virtue of the archive risk use case that we focus on here is its simplicity, as it provides a straightforward anchor to the complexity of trying to teach technology to simulate human judgment for content processing.

1. | When Risk Emerges from Within

1.1. | CHANGING RISK CAN EXPOSE VULNERABILITIES

The pace of change in business and technology is grueling. Even if companies deal proactively with change, managing archives tended to land at the end of the priority queue because the cost of purging post-retention documents was higher than the risk. However, with heightened risk, companies now face exorbitant costs as a result of data breaches, as well as brand and reputation damages, so the cost-benefit ratio has changed dramatically.

Filing cabinets have mostly been retired, and the digital replacement documents are often scans or other digital formats, such as PDFs. They're usually buried in packaged software or arcane directory structures, leaving document management systems (DMSs) to become the new digital basement where things are stored and long forgotten. These systems house everything from patient records and insurance policies to mortgage files, legal documents, and background or credit checks.

Changing technologies, acquisitions, organizational processes and regulatory requirements may complicate the archive purging process and force significant manual activity as workarounds. Entropy therefore ensues. So, where do many companies stand today?

- Most companies have multiple archives from acquisitions, resulting in decentralization or disparate systems
- Some DMSs may not be formally defined as archives, but still hold information that could be purged
- Sensitive documents are likely to vary in form and content for many reasons (internal or external)
- Documents may be scattered, duplicated, appended and randomly replicated (e.g. email). They may be combined in obscure ways, disconnected or missing information
- Critical risk documents may be buried in piles of unrelated content material

1.2. | CONTENT PROCESSING TECHNOLOGY CAN HELP

To make sense of the chaos, organizations should consider implementing a content processing pipeline. Leveraging a content processing solution can:

- Provide an automated pipeline that can process documents, images, audio and video
- Extract key information for companies to leverage as insights
- Interpret and understand words, phrases, sentences and clauses
- Normalize language variation, format, metadata and indexing
- Apply tags and linkage to enrich content using taxonomy or semantic markup
- Flag exceptions or uncertain items
- Store a pure, enriched version of the content to use later in other applications and analysis, increasing ROI
- Purge obsolete content to follow compliance standards (after appropriate notification and review)

However, getting to meaningful accuracy and cost benefit takes some discipline. An initial proof of concept (PoC) for a content processing solution may dazzle and show promise but fail to scale to the full archive scope. It may require a long tail of optimization, federated techniques, multiple processing paths and curation. An experienced vendor can shorten the journey and de-risk the investment while transferring content processing capability to in-house teams.

1.2.1. ON ACCURACY: HOW RIGHT IS RIGHT ENOUGH?

Content processing platforms teach a computer to make decisions like a human. Unfortunately, with many artificial intelligence (AI) PoCs, the question of how accurate you need to be is typically an afterthought. Accuracy is not a colloquial term in this domain. It has a very specific formula based on two other terms: Precision and recall. Precision is the percentage of your identified set versus what you were supposed to identify and recall is the correctness within the identified set itself. Your 'accuracy' is a combination of the two.

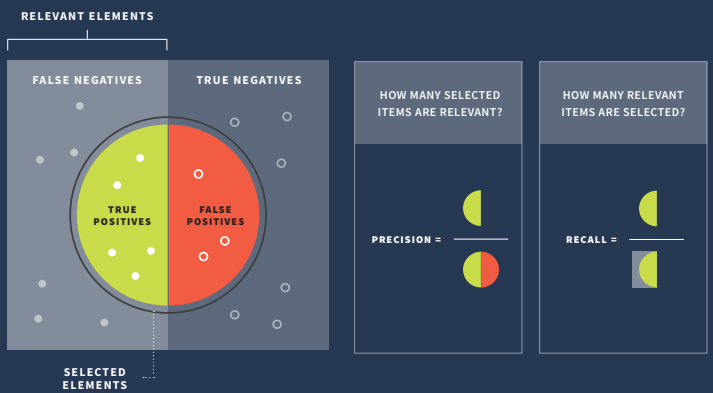

FIGURE 1: ACCURACY, PRECISION & RECALL

Getting AI to duplicate the intelligence of a professional with decades of knowledge and experience is hard, and costs rise drastically to achieve between 70-90% accuracy. Achieving 100% accuracy is impossible (except for simple processes such as binary sentiment). For that reason, it's critical for people to work alongside the content flowing through the pipeline from the beginning to correct mistakes, enhance the knowledge base as new content, concepts, terms and phrases come through, and maintain the training, tests and other tasks that optimize the platform. This role may be called curator, steward, exception handler, trainer, supervisor, librarian, knowledge staff—we know them as 'humans in the loop.' This role complements real intelligence with the artificial type.

2. | A Reference Model for Content Pipelines

The comprehensive content pipeline reference model below serves as a starting point to map out what your organization may need for implementation. There are many options for how to implement the solution depending on your current in-house IT strategies, tooling and components.

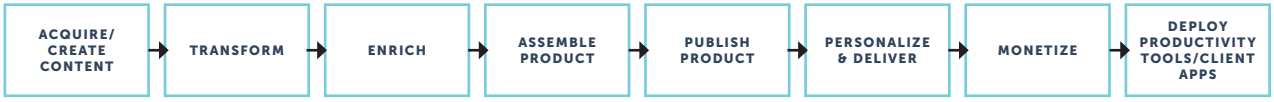

FIGURE 2: CONTENT FRAMEWORK

This model catalogues content processing functions through a multi-level framework that drills down to individual technology services and business activities. It's technology agnostic, but assures full coverage when used to map a subset of functions for a particular application. Additionally, the model provides a system architecture baseline and a roadmap that runs through new application areas, curbing the incremental costs of added investment.

2.1. | LEVERAGING A PII ARCHIVE CLEAN-UP APPLICATION

From the reference model, you'll need to select the appropriate content processing application functions and tailor them to any given use case (archives, for the sake of this paper). In doing so, you could end up with the following framework:

FIGURE 3: GENERAL REPOSITORY CLEANUP FUNCTIONAL MODEL

It all seems easy enough, but it's crucial to remember that, while computers are great with numbers, language processing can be a major challenge—even slight content variations will throw off the best natural language processing (NLP) or machine learning (ML) solutions. The more complex the concepts, the harder it is for the technology to understand.

It takes a truly agile approach to be successful, which involves swapping components and adding configurations based upon new samples or exceptions. You are likely to end up with federated components — some performing similar functions, but with different techniques, algorithms or feature emphasis. For example, not all optical character recognition (OCR) solutions are the same. Some are better at identifying regions, some are better at damaged hardcopy images and others are better at processing handwriting or unusual fonts. It's a process, and there are certainly advantages to being able to draw on deep expertise for advice.

2.2. | ARCHITECTURAL THINKING

While some functions in this domain, such as work flow, are standard, some unique architectural thinking will also be required to accommodate many more edge cases than with other processing types.

Here are some core functions that you need to accommodate:

- Workflow, including roles, queue allocation, triage, routing, approvals, exceptions, escalation, supervision, shifts, holds, notes, collaboration and reporting
- Knowledge base (for people and computers), including training and accuracy testing of sample sets
- OCR, extraction conversion and enrichment (NLP/NER)
- UI for curation, exception processing and training set management, and business approval
- Disposition (approval, target repository, destruction, compliance reporting, agnostic, anonymized)
- A place to save normalized, enriched content without loss of fidelity due to output formatting
- Traditional industry-specific indexing schemes and output processing

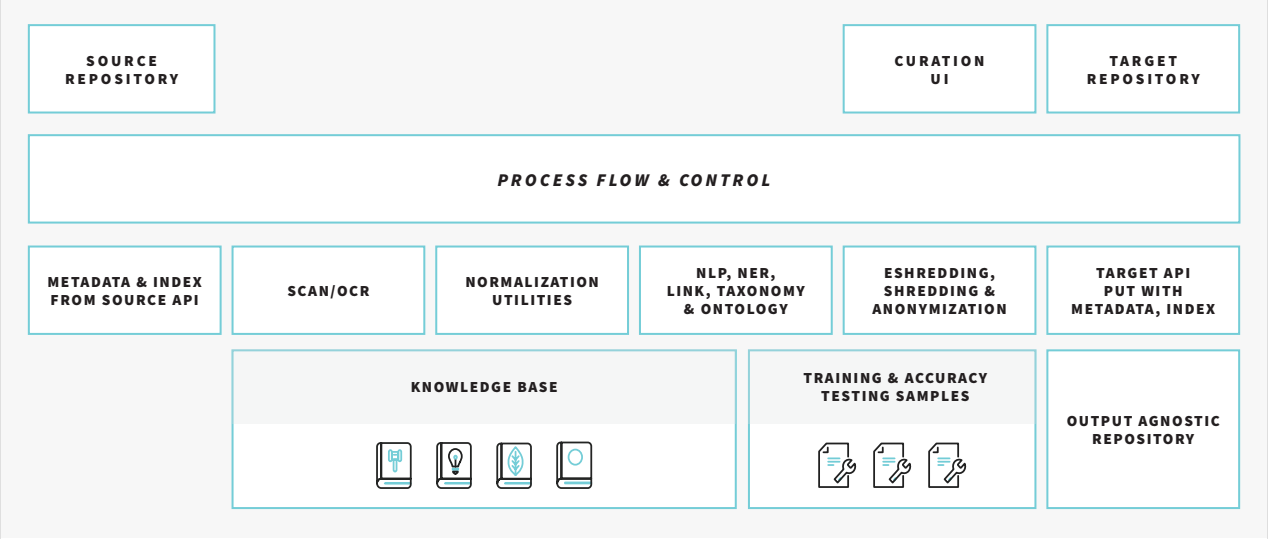

FIGURE 4: SAMPLE TECHNICAL ARCHITECTURE

This is not a domain where standardizing leads to success. It's important to consult your InfoSec and audit officers early in implementation to architect a content processing solution that complies with organizational and regulatory standards.

3. | Choosing the Most Effective Tech for Your Business

Fortunately, there are many options available to develop the right technology ecosystem for a comprehensive content processing solution—so many, in fact, that it can be overwhelming. However, your enterprise culture and requirements may help narrow down options.

One topic to consider among these choices is the maturity level of certain technologies. For example, AI, NLP and ML are complex tools with long optimization tails and declining cost benefit. Newer functionality in this space may not provide the accuracy and breadth of features of more mature components. Let's examine a few high-level categories:

3.1. | TRADITIONAL DMS VENDORS

Traditional DMS vendors may provide some features that will address your needs and they may be good enough to get the job done, but using a traditional DMS vendor means that you'll rely on more manual work from curators to process and check for more exceptions. You might want to augment the architecture of the platform suite with individual commercial off-the-shelf software (COTS) or open source components for more capabilities and reduced manual effort.

3.2. | OPEN SOURCE SOLUTIONS

Open source content processing solutions are typically mature, as they've undergone decades of development. In this domain, open source is often equal to or better than commercial software on a functionality basis. There is perceived bias that total cost of ownership is higher due to low-level code integration. However, this can be offset against upgrade cycles and service fees. It can also be mitigated by a capable vendor partner that may bring accelerators to the task. On the other hand, some companies have enterprise policies that dictate a support contract around open source from one of the wrapper service providers. These procedures can escalate 'free' open source up to six-digit support contracts, reducing the upfront advantage of low costs. Even so, the performance of some open source content components can be compelling.

3.3. | COMMERCIAL OFF-THE-SHELF (COTS) SOFTWARE

Good vendor partners can combine COTS and open source capabilities, or stick strictly with COTS if that's your company's strategy. COTS solutions typically require less integration and may have more user-friendly interfaces. These tools may also provide starter models for common business applications, and vendors will offer consulting and support services.

Examples include Cogito and Adobe on the broader side, and ABBYY or Nuance in the OCR/normalization space. Others, like Net Owl and LexMachina, are a bit specialized to target named entity recognition (NER) and legal content, respectively. PoolParty and TopBraid are examples of knowledge repositories. Workflow vendors like Pega and Appian have also added cognitive components.

3.4. | ALL-IN-ONE COGNITIVE PIPELINE PLATFORM VENDORS

The latest entrants to the content processing domain provide a unified, 'out-of-the-box' cognitive platform. These solutions are a great option for simple- to medium-complexity use cases, such as data extraction. They also accommodate functionality for more complex or edge cases, including archive document content processing.

These platforms fly under the banners of robotic process automation (RPA), cognitive process automation (CPA), cognitive document processing (CDP) and intelligent automation (IA). Examples include WorkFusion, BluePrism and UiPath, which provide a fully integrated set of features covering everything we have discussed above, including curation. They come with tools and training, as well as a consulting ecosystem, to help you get started quickly and branch out to other use cases.

3.5. | CLOUD VENDORS

It's worth mentioning cloud vendors, as most major cloud providers now offer tools, components and accelerators for content processing to make the core service more compelling. Some of these components are derived from open source equivalents and supply similar features and capabilities. You may find these components are tuned to the more common use cases, and may be 'black boxed'—removing some configurability. But they can be a great starter set that you can augment if you require more accuracy and capabilities or greater optimization for more complex work.

3.6. | 'AS A SERVICE' BPO

Lastly, some vendors provide the entire content processing solution, including curation as a service. This can be an ideal option in cases where you don't permanently need this solution, or want to avoid capital expense. Complete service providers will typically follow a method that resembles this:

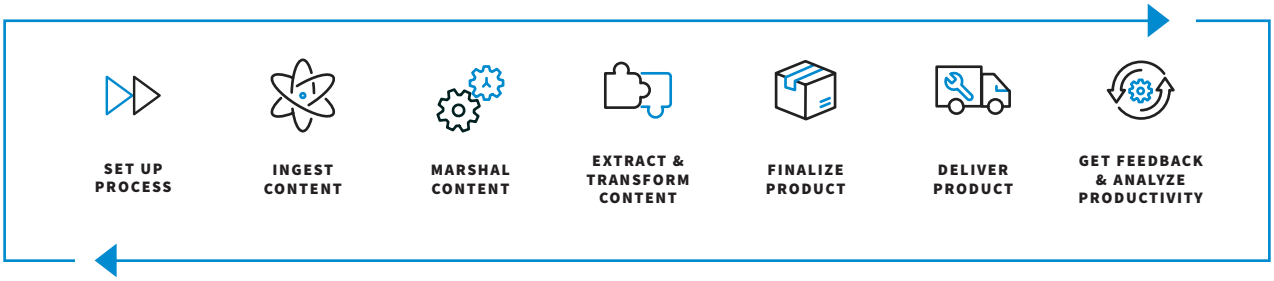

FIGURE 4: SAMPLE CURATORIAL METHODOLOGY

3.7. | PLAN FOR A SMALL DEPARTMENTAL TECHNOLOGY BUDGET

AI solutions in general run on a curve with a long tail. You'll want to provide your operations teams with some incremental technology support to make continual, small improvements. This may include enhancements to existing components or configurations of known types of content, which are a result of different types of content and edge cases that didn't appear in an initial analysis, or new content types, formats, requirements and changes in retention rules.

4. | Conclusion

We've now defined a functional approach to using a content processing pipeline to address GRC for the archive document use case. By considering the sometimes overlooked case of PII in purgeable archives and exploring the many options available from both an approach and technology perspective based on the needs of your specific organization, you will have the tools and insight to begin to roadmap against your current technology portfolio and internal capability for curation.

In closing, we would like to offer a sample menu of additional use cases that will fit right into this domain:

- ALLOWING FOR ADVANCED 'UNSTRUCTURED' CONTENT PROCESSING IN YOUR BIG DATA ENVIRONMENT
- PROCESSING VARIATIONS IN LEGAL CLAUSES ACROSS JURISDICTIONS AND REGULATORS, AS WELL AS MATCHING FORM CLAUSES AND SCHEDULES
- IDENTIFYING EXPOSURE TO EMERGING RISKS THAT HAVE ARISEN FROM IN FORCE CONTRACTS, POLICIES OR OTHER TRANSACTIONS DOCUMENTED IN YOUR ARCHIVE
- UTILIZING SOCIAL MEDIA FOR MARKETING OR IDENTIFYING POPULATIONS (E.G. CLINICAL TRIAL CANDIDATES)
- PROCESSING CLAIMS INVOLVING INVOLVING FRAUD DETECTION, DAMAGE IDENTIFICATION, REPAIR ESTIMATION, OR TOUCHLESS PROCESSING
- MONITORING THE COMPETITIVE INTELLIGENCE OF OTHER INDUSTRY PLAYERS, REGULATORS AND LEGISLATORS
- PROSPECTING SALES (E.G. TRACKING PROSPECTIVE B2B BUYERS BY MONITORING EXECUTIVE TURNOVER, NEW STRATEGY OR EMPLOYEE TURNOVER)
- CONNECTING CUSTOMERS, HOUSEHOLDS AND GROUPS ACROSS LINES OF BUSINESS FOR MARKETING, SALES OR RISK CONCENTRATIONS
- TRAINING CONVERSATIONAL AGENTS (OR CHATBOTS), FOR CUSTOMER SERVICE OR ECOMMERCE
- EVALUATING CUSTOMER SERVICE QUALITY AND COMPLIANCE THROUGH AUDIO INTERPRETATION
- SCANNING THE INTERNET FOR RISKS ASSOCIATED WITH SPECIFIC GEOGRAPHIES, ASSOCIATIONS AND EVENTS
- CONNECTING SOURCES OF INFORMATION AROUND CATASTROPHIC SITUATIONS
- MONITORING GEOGRAPHIC MEDICAL, HEALTH AND ENVIRONMENTAL RISKS
- BRINGING QUALITATIVE BACKGROUND INFORMATION TO AUGMENT QUANTITATIVE RISK EVALUATION
- ALLOWING INLINE FEEDBACK TO CUSTOMERS TO CUSTOMERS SENDING PHOTOS FROM THEIR PHONE
- INTAKING DOCUMENTS FROM CUSTOMERS AND COMPANIES IN CASES WHERE TRANSACTIONS AND PROPOSALS START WITH A LARGE PACKAGE OF INFORMATION AS A BASELINE

Content processing technology can be applied to many different use cases. The best approach to content processing in any of these cases is tackling it holistically. If you already have some capabilities in place, the PII use case we've shared in this white paper may lend insight into how you can leverage content processing elsewhere. Once you've established your solution, you'll have a repeatable method to mitigate risk, reduce cost and increase performance.

ABOUT EPAM

Since 1993, EPAM Systems, Inc. (NYSE: EPAM) has leveraged its software engineering expertise to become a leading global product development, digital platform engineering, and top digital and product design agency. Through its 'Engineering DNA' and innovative strategy, consulting, and design capabilities, EPAM works in collaboration with its customers to deliver next-gen solutions that turn complex business challenges into real business outcomes. EPAM's global teams serve customers in more than 30 countries across North America, Europe, Asia and Australia. As a recognized market leader in multiple categories among top global independent research agencies, EPAM was one of only four technology companies to appear on Forbes 25 Fastest Growing Public Tech Companies list every year of publication since 2013 and was the only IT services company featured on Fortune's 100 Fastest-Growing Companies list of 2019.

Learn more at **ww.epam.com** and follow us on **Twitter @EPAMSYSTEMS** and **LinkedIn.**

GLOBAL

41 University Drive,
Suite 202
Newtown, PA 18940, USA

P: +1-267-759-9000

F: +1-267-759-8989

B829_v1_09/2020

<epam>