WHITE PAPER

Content Lakes: A Language-Centric Content Storage & Processing Solution

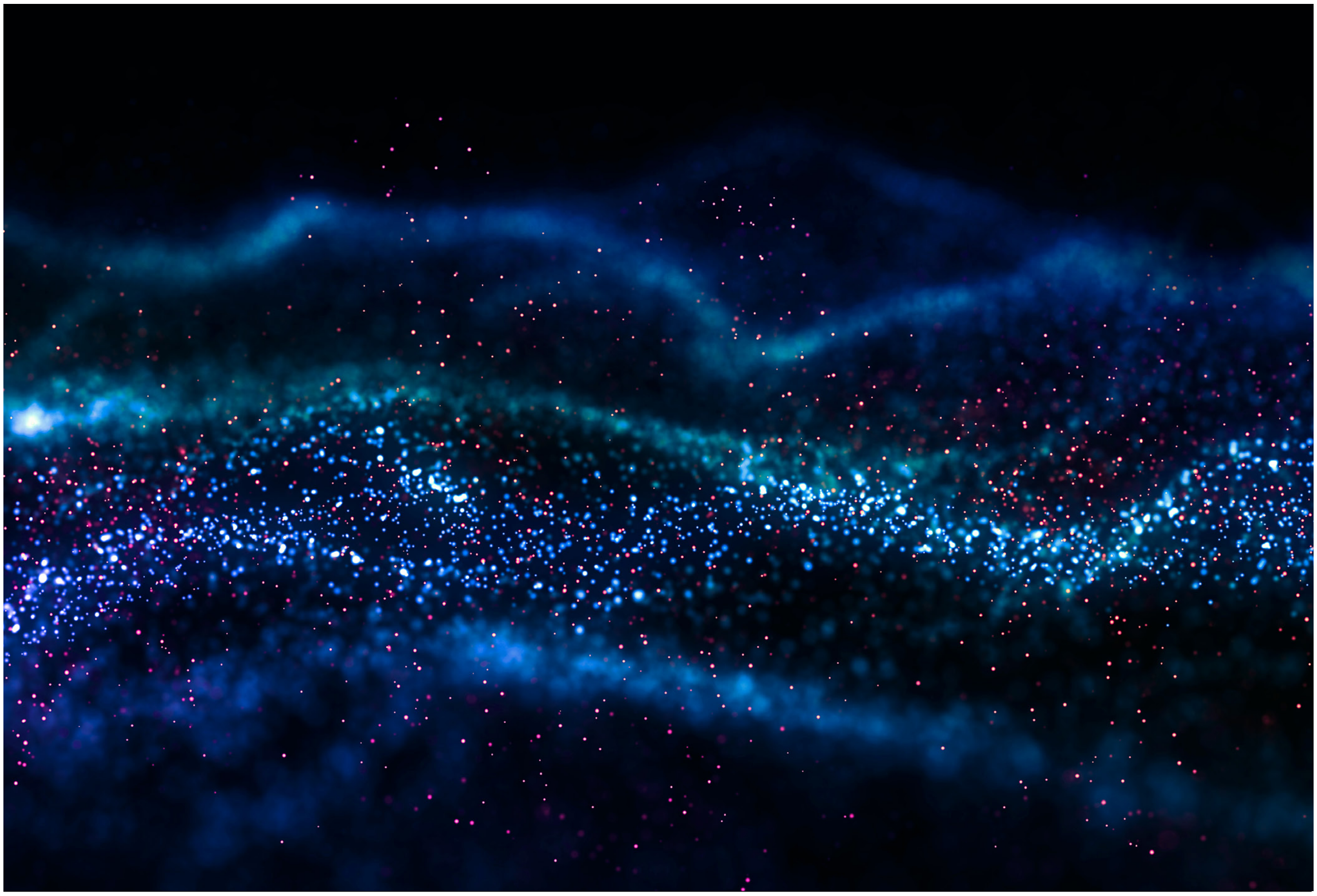
Language-Centric, Content Insights

Contents

A Brief Introduction to Data Lakes

It's the age of analytics and data insights have become a critical part of every business. Every year, enterprises amass terabytes upon terabytes of data to extract relevant information and analyze it for valuable patterns. These large quantities of raw, unfiltered, structured, semi-structured and unstructured data are stored in data lake repositories.

Oftentimes, data lakes are a component of big data platforms, mitigating the need for immediate processing, archiving or deleting. Without having to go through costly collection activities, data lakes allow for additional derivation and completely new analytical spokes. While one of the core principles of a data lake is that it stores and processes raw data for organizations to access and organize at any point, they're not optimal for language or unstructured content.



The Rise of Content Lakes

From scientific research papers and legal documents to mortgage information and insurance claims, 80% of all enterprise data is considered unstructured. This surge in digital documents and communications, otherwise known as Big Content, has completely shifted the paradigm of data collection and categorization because this information is key to unlocking highly personalized experiences and extremely precise results to meet customer demands.

Many would argue that most "unstructured data" isn't unstructured at all. In fact, human language is a highly structured, formulaic expression of concepts, ideas and emotions. For those who employ Big Content, language is primary and data is derivative.
Just like Big Data, Big Content needs a strategy for enhancing and refining data and, of course, a raw data repository conducive to human language.

That's where content lakes can come into play. A content lake is a specialized form of data lake that stores language-based content and offers broad-use access to original content as well as analytics, tabular data derivation and statistical processing. Unlike data lakes, content lakes assume that language artifacts are highly structured and that users need to access the original content as a whole, including both language and data insights.

Potential Content Lake Use Cases

- Insurance policy claims & actuarial data
- Legal research & predictive analysis
- Patent IP crawling
- Scientific research
- Healthcare treatment, research & diagnostics
- Trading indicators
- Competitive monitoring, research & analysis
- eCommerce cross-selling & advertising yield
- Media marketing analysis & one-to-one marketing
- Customer service
- Ad placement engine modelling
- Brand value monitoring
- Physical risk forecasting
- Cybercrime, fraud, terrorism & suspect/target identification
- National security
- Translation
- Conversational agents
- Knowledge management
- Emerging trend alerts

The Rise of Content Lakes (cont'd)

Understanding when to Employ a Content Lake

When the majority of your content, whether produced or acquired, includes human language expression, it might be time to implement a content lake architecture. For example, customers who pay for content, such as research papers, legal documents or patents, need access to the original piece of content as a whole. Even in some edge cases, like a government organization that analyzes large amounts of voice conversations and ties them to smaller databases, a content lake would be extremely useful for storing and organizing original content.

Here are some indicators that you should implement a content lake architecture and techniques:

- **Language-based content:** Your "data" is mostly human language-based content, whether written, graphic, audio, video, etc.
- **High cost of acquiring & maintaining content:** You pay a lot for content and may have similar content from multiple sources
- **Meaning vs. data:** You need more from the content than just extracted data, including recognition of concepts and context, as well as a comparison to other relative content
- **Original content:** You or your organization produce content that people or institutions purchase
- **Complex ideas:** Your use of the content is complex, and you need the original content to be accurate, complete and timely
- **All the facts:** You or your customers need to know that all of your content is accurate and current; this includes the ability to identify subtle distinctions between revisions and duplicates, as well as corrections and amendments
- **Original access:** You or your customers need to see the original source, or raw content

Tackling Pre-Normalization to Reduce Processing Costs

One of the core data lake principles is having raw data available anytime, anywhere, so that end users can easily go back and analyze the original source.

But as organizations transition from print to digital, many incoming materials are hard copies, and going all the way back to the original source can be costly. This raw content must be prepared so that end users can access relevant information quickly and easily. With human language, some initial normalization, or the process of structuring to make content machine readable, helps improve storage and processing dramatically.

Examples of normalization include audio-to-text and spatial recognition of image and video attributes. For paginated sources you may need decoding, fragment re-composition and addendum tokenization, as well as attributes like page, chapter and outline numbers; tokenized notes and citations; hand-written marginalia; scope of indices and references; and any source mark-up.

In a content lake, normalization can be expensive to leverage broadly, which is why it's essential for organizations to employ a dedicated pre-normalized repository. Pre-normalizing materials involves getting content into its base schema and prepped for normalization. Converting sources to a particular schema or style eliminates the need to navigate back over and over. That way, organizations can leave source materials in their raw, digital state for systems to look at labels and pull relevant information accordingly.

The more diverse the uses of the content, the more likely a pre-normalized repository will help alongside normalized and master repositories. In fact, if organizations require multiple masters, or have legacy investments that require a hub-and-spoke approach, the pre-normalized repository is absolutely a pre-requisite for a successful content lake.

The Rise of Content Lakes (cont'd)

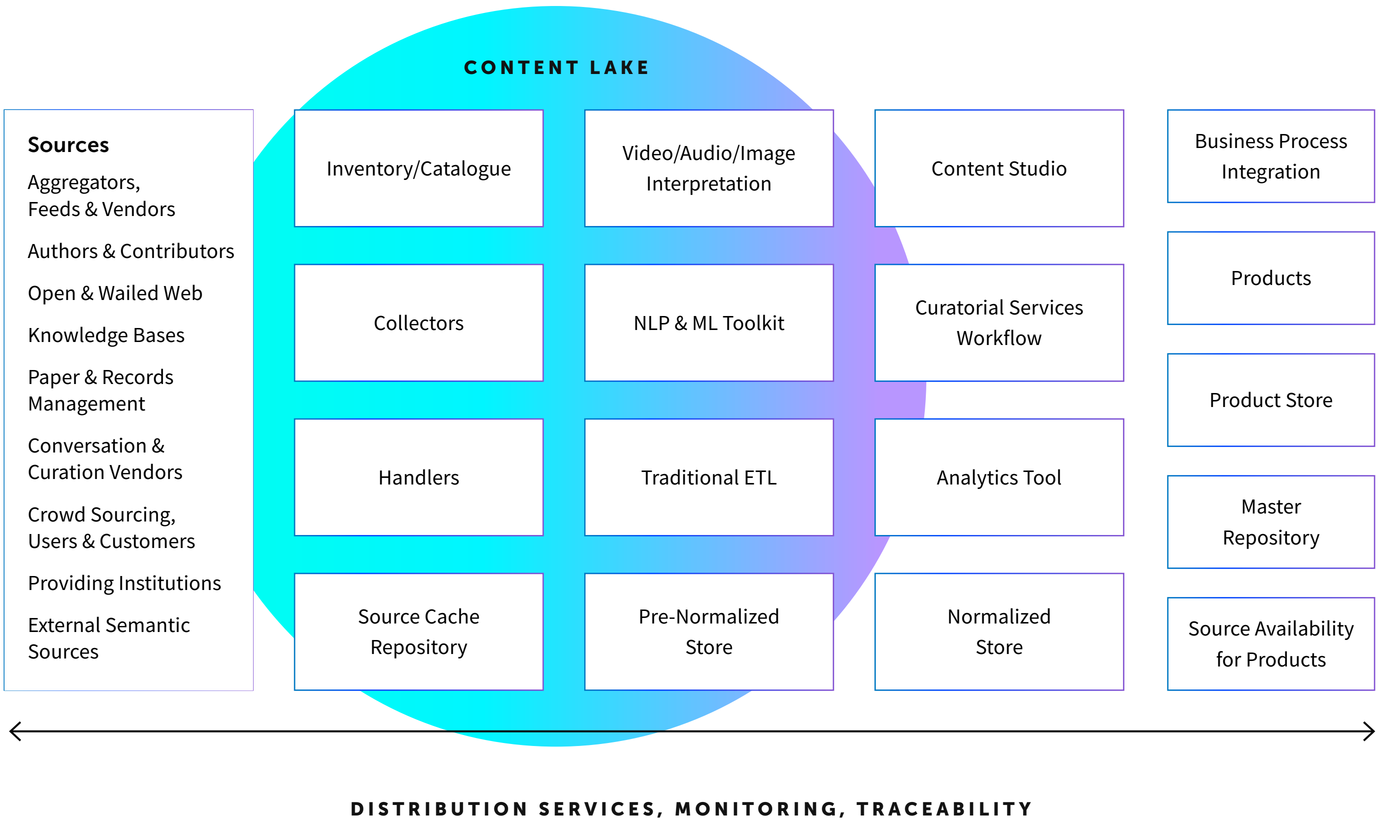
Other Content Lake Best Practices

Pre-normalization and normalization are only a couple of best practices to consider for content lakes. Aside from these, there are several other tools and techniques that are necessary for successful implementation:

- An inventory or catalog to assure that content is accurate and easy for end users to locate
- Alerts for when content is not accurate, along with specialized ingestion functionality to remediate
- A platform that understands human language, which often involves:
    - Investment in tools and capabilities for human language processing, which will be significantly larger than for data extraction, cleansing and aggregate calculations
    - Natural Language Processing (NLP) toolkits, enrichment systems, taxonomy/ontology, Named Entity Recognition (NER) and relationship recognition, such as formal citations and computational linguistics
    - Federated ML/AI/NLP for high accuracy enrichment where machine learning, symbolic language processing and artificial intelligence can be deployed to human intelligence assistance
- Access to original artifacts, such as documents, media, physical records, etc.
- Authoring and curatorial tools, as well as an editorial platform
- A curatorial pipeline, workflow and semantic mark-up tools to top off ML and symbolic-based AI to ensure high accuracy of content for end users who pay top dollar
- Handler components for physical records, like paper, binders, books and tapes, such as OCR, audio-to-text, visual recognition, graphic decomposition, raster to vector and other media conversion tools
- Caching and version comparison for content accuracy as well as understanding when a crawled source site has changed format

The Rise of Content Lakes (cont'd)

A Service Framework for a Content Lake



CONTENT LAKE

Sources

- Aggregators, Feeds & Vendors
- Authors & Contributors
- Open & Wailed Web
- Knowledge Bases
- Paper & Records Management
- Conversation & Curation Vendors
- Crowd Sourcing, Users & Customers
- Providing Institutions
- External Semantic Sources

Inventory/Catalogue

Collectors

Handlers

Source Cache Repository

Video/Audio/Image Interpretation

NLP & ML Toolkit

Traditional ETL

Pre-Normalized Store

Content Studio

Curatorial Services Workflow

Analytics Tool

Normalized Store

Business Process Integration

Products

Product Store

Master Repository

Source Availability for Products

DISTRIBUTION SERVICES, MONITORING, TRACEABILITY

Fast Forward: The Future is in the Cloud

While storage is “cheap” today, it’s not free. For organizations that have successfully tackled pre-normalization of aggregated content, there are next-generation tools that provide foundational components for services, but configuring the logic for unique sources can be costly. Modern architecture and design, as well as a scalable cloud infrastructure and network stacks, can offer significant cost savings to offset the investment associated with large-scale content and data storage.

Horizontal scaling lets organizations deal with variable load by adding dedicated server capacity in minutes or hours and virtual server capacity on demand. Using continuous integration techniques along with the cloud, organizations can also get unprecedented environment management, including spin-up and shutdown.

Organizations that are the sole consumer of content (i.e., the content is just being used internally) can conserve and avoid long lead times with bare-metal captive data center infrastructure by spinning up redundant environments, testing, sandboxes, processing streams, analytics development and mart consumption environments quickly by using an agile schedule.

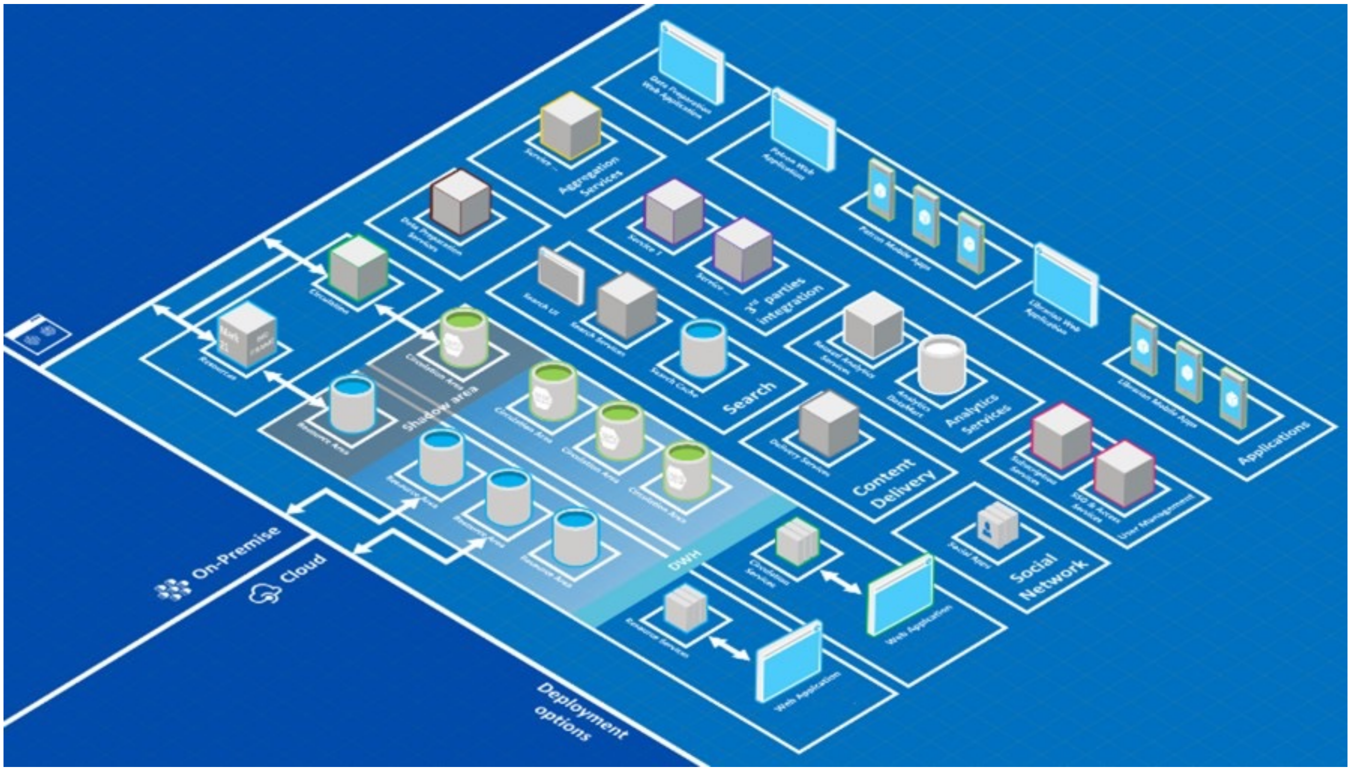
Those that use content as a revenue source can also benefit from current technology and techniques. For example, companies with advertising revenue alongside a platform or service, companies with high-value content products and companies that rely on subscription fees for access to content will gain tremendous value from contemporary customer-centric, lean, agile and continuous product management.

With the ability to easily maintain conceptual integrity, organizations can also avoid content silos. To help maintain conceptual integrity, modern and flexible schema techniques provide conceptual integrity across various content types and sparse metadata/taxonomy/ontology often found in different sources. Companies should use flexible schema to avoid brittle processing streams, while controlling an element catalog against ongoing ontology and taxonomy disciplines. This both reduces maintenance cost and eases upgrade complexity when refactoring legacy source acquisition down the road.

Aside from shortening time-to-market and connecting customers directly to your product development capabilities, environment management and scalability for specific product-based content processing becomes significantly more efficient, flexible and direct, at high velocity through the cloud and other methodologies. Meanwhile, rapidly scalable testing and curatorial content services mean interoperability testing can fit into agile cycles rather than seasonal or annual cycles.

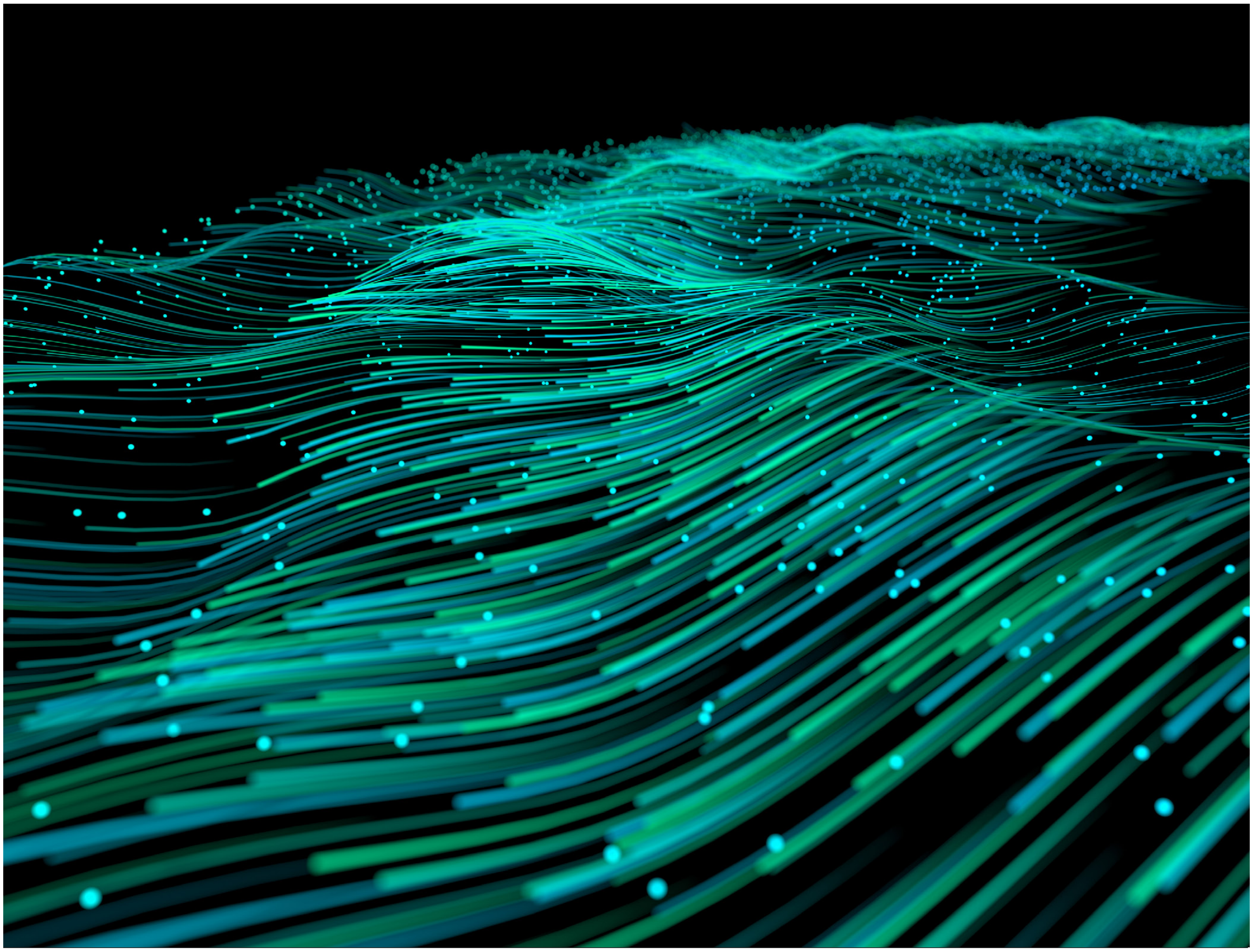
Fast Forward: The Future is in the Cloud (cont'd)

Scalable Cloud Architecture & Network Stacks



Facing the Digital Age with Confidence

Many organizations that handle large quantities of unstructured human language content have been feeling the full weight of disruption for some time. Now, content producers and purchasers alike are being challenged to not just embrace the migration to digital but also leverage emerging technologies to deliver optimized experiences both internally and externally. Content lakes are an extremely valuable method that, when implemented correctly, can completely upend traditional content processing by allowing end users to automatically reach relevant, original sources vital to their business.



Are you interested in learning more about how implementing a content lake could improve your business? **Contact:**

James Buonocore

TITLE

Solutions Consultant, EPAM

CONTACT

James_Buonocore1@epam.com

Oleg Vilchinski

TITLE

VP, Global Head of Business Information Solutions, EPAM

CONTACT

Oleg_Vilchinski@epam.com

About EPAM

Since 1993, EPAM Systems, Inc. (NYSE: EPAM) has leveraged its advanced software engineering heritage to become the foremost global digital transformation services provider – leading the industry in digital and physical product development and digital platform engineering services.

Through its innovative strategy; integrated advisory, consulting, and design capabilities; and unique 'Engineering DNA,' EPAM's globally deployed hybrid teams help make the future real for clients and communities around the world by powering better enterprise, education and health platforms that connect people, optimize experiences, and improve people's lives. In 2021, EPAM was added to the S&P 500 and included among the list of Forbes Global 2000 companies.

Selected by Newsweek as a 2021, 2022 and 2023 Most Loved Workplace, EPAM's global multidisciplinary teams serve customers in more than 50 countries across six continents. As a recognized leader, EPAM is listed among the top 15 companies in Information Technology Services on the Fortune 1000 and ranked four times as the top IT services company on Fortune's 100 Fastest Growing Companies list. EPAM is also listed among Ad Age's top 25 World's Largest Agency Companies for three consecutive years, and Consulting Magazine named EPAM Continuum a top 20 Fastest Growing Firm

Learn more at 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