milestone technologies

WHITEPAPER

The Industry Blueprint for Capturing GenAI Value in 2025

Explore how various industries are strategically preparing for the future with GenAI in action, driving innovation and competitive advantage.

CONTENTS

Background

“Data is everything” and "Everything is data" are foundational beliefs that mirror each other in the realm of AI. Today, we’re capturing every bit of data that matters and transforming it into intelligence that the world is using to make informed decisions and generate value in ways previously unimagined.

As generative AI goes mainstream this year and permeates almost every aspect of business and social life, the breakneck speed of data generation is taking the world by storm. For example, a high-end connected car is generating close to 380GB of data per hour, and by 2030, it’s estimated that 95% of all new vehicles produced globally will be connected. NISAR, a satellite developed jointly by NASA and ISRO, and expected to launch later this year, is projected to generate 85TB of data on an average per day. CERN, the largest particle physics laboratory in the world generates 1 petabyte (that’s 1000TB) of data daily. And ChatGPT generates an undisclosed but arguably stupendous volume of data each day – responding to an estimated 10 million queries.

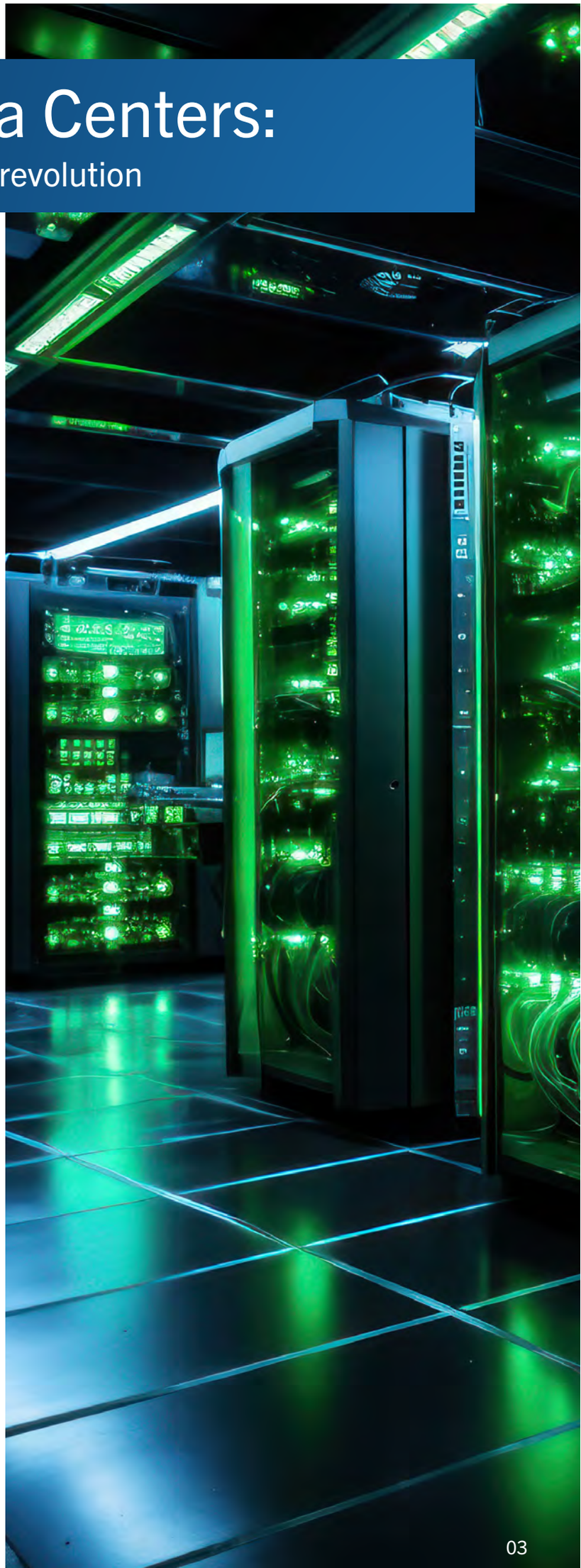
All this is but a preview of the quantum of data that GenAI creates. Experts predict that by 2032, GenAI will balloon into a $1.3 trillion revenue industry, with the global datasphere doubling within 9-12 months, as against the current 24-month cycle.

That said, the IT reality of today – consisting of mainframes, previous generation architectures, and legacy systems, is not geared for the explosive growth and evolution of GenAI. This leaves us with the following questions: What kind of infrastructure is needed to bridge the modern-day gap and prepare for the constantly growing GenAI workloads? How are GenAI practitioners gearing up for the current and next wave of computational demand? And what strategies are businesses adopting to unlock the full potential of GenAI?

Broadly-speaking, there are four tracks along which these practitioners will battle for AI leadership.

TRACK 1- Data Centers:

The new frontiers for GenAI revolution



Data centers, including hyperscalers, are the unsung heroes of our digital world, working behind the scenes and powering our daily operations – be it mobile banking, social media activity, online gaming, cab hailing, or generating content. While GenAI is as promising as the new dawn, it's also as ravenous as a beast – blame it on its insatiable appetite for computational and electrical power and the demand it makes on storage. At the back end, it takes no less than the robust infrastructure of a data center to do all the heavy lifting needed for processing GenAI-intense workloads.

But as of today, the current data center infrastructure is wanting. Moore's Law has run its course and general-purpose chips or CPUs cannot support high performance computing (HPC) demands – simply because the transistor size and chip density have reached their practical and physical limits. Semiconductor companies such as NVIDIA and hyperscalers such as Meta, Google, Amazon, Microsoft, and Apple have acknowledged this and are developing AI chips – specialized integrated circuits that use parallel processing to perform multiple tasks with reduced latency.

These marvels are designed to accomplish greater accuracy and energy efficiency – significantly outperforming conventional CPUs that undertake sequential processing. A noteworthy feature is that the computational power of AI chips doubles in just about six months – a stark contrast to the two-year window applicable under Moore's law. Researchers are also experimenting with multi-die designs (chiplets with multiple chips) stacked for achieving greater density and compute power, and this promises to be another game changer.

Computational Power of AI chips

As such, traditional data centers are evolving into AI data centers with ultra-modern chipsets, processors, memory kits, and graphics cards. Processors are being reimagined, with the x86 standard making way for the faster and more efficient open-source RISC-V or the proprietary ARM architectures for supporting AI and ML workloads. In effect, ARM will do to x86 what x86 did to mainframes. This is evident in the fact that the new Mac computer, the new Microsoft Operating System, Microsoft Azure Cloud, and AWS Cloud are embracing ARM standards.

New Architectures for Processors

Unlike the x86 predecessor, ARM delivers parallel computing, favors simplicity and speed, produces less heat, and performs processing, inference, and postprocessing on a single piece of silicon. RISC-V, on the other hand, outperforms ARM in the openness, customization, innovation, and cost-efficiency it offers in low-power processor core designs.

AI-intense workloads also require advanced GPUs. More recently, Meta announced its plan to build a massive computing infrastructure by acquiring 350,000 Nvidia H100 GPUs to augment their AI capabilities. Today, some of the high-performance GPUs are designed to reach up to 1,979 TFLOPS of FP16 performance.

This year could also witness a memory solution like never before – following a recent technological breakthrough. We're talking about devices that boast DRAM latency, non-volatile memory, and ultra-low power consumption – all rolled into one. They may make their way into commercial production sooner than imagined. This innovation will use phase-change memory and may arrive in a 3D-stacked form factor.

When run on such non-volatile DRAM memory, neuromorphic computing systems, which are modeled on human brain architecture, can catalyze groundbreaking advancements in hyper-realistic generative models.

The extraordinary compute needed for meeting the soaring demand for GenAI can only be provided when an equally dizzying quantum of electric power is available. To put things in perspective, the energy drawn by ChatGPT on a single day could easily power 33,000 homes. AI data centers, including hyperscalers, are rising to the energy challenge by increasing the electrical output per rack. This is creating an increasingly dense and sophisticated data storage and computational infrastructure.

While the densities of data centers were in the range of 3-7kW (or at most 10-15kW) per cabinet per rack for traditional AI support, the same is now exceeding 100kW with GenAI taking center stage. High-Performance Computing (HPC) data centers and hyperscalers are now signing power leases in the range of 100-500 MW, and talks are also on for building gigawatt capacities in the immediate short term, with liquid cooling being the norm for thermal management.

For rapid, scalable data center expansion, enterprises can choose prefabricated components that can be stacked onto their existing infrastructure, or they can opt for a fixed-size prefabricated containerized unit that can be set up outdoors – both options providing the cost and energy efficiency of modular data centers. The complete hardware revamp at data centers is projected to drive the largest capital expenditure cycle over the next decade, with spending projected to reach $400 billion by 2027.

Capital Expenditure Cycle by 2027

For the set-up, relocation, consolidation, or day-to-day functioning of any data center, the requirement of technical knowledge, expertise, observability, analytics, and ongoing support cannot be overstated. Data centers need dedicated teams – across engineering, development, IT Ops, data security, risk, and compliance functions. Being one of the greatest powerhouses of a nation's economy, there's a lot riding on them, so it's not just about keeping the lights on. There are several performance aspects that need to be considered – energy and cooling efficiency, service availability, server utilization, break-fix maintenance, and Asset Lifecycle Management (ALM) are just a few of them.

With a strategic partner, not only are the risks of downtime and costly errors avoided, but the data center also accomplishes greater operational efficiency, ROI, security, stability, scalability, effective utilization of assets, and much more.

Milestone has been an IT consulting and services partner to businesses for over 25 years and also drives value for eight of the top ten tech companies in the U.S.A. Its profound understanding of data centers, coupled with deep expertise in cloud, AI, automation, big data, analytics, networking, and data center operations, positions it as a pivotal partner and key enabler for growth. Milestone's Integrated Operations Center offers 24/7 monitoring of IT infrastructure and applications, and enhances operational efficiency, minimize response time, keeps systems running smoothly, and provides observability into the IT environment. With expertise in data center hardware, software and network infrastructure, Milestone's services are aimed at providing tailored services, end-to-end deployment, reliability, scalability, and day-to-day break-fix maintenance support – helping data centers phase out legacy systems and dated hardware and become AI-ready. Recently, Milestone successfully relocated 250000 servers, replaced 500000 HDDs, and audited 300000 assets for several data centers. For another data center, Milestone's 'pizza box' project amped up its computing power six weeks ahead of schedule.

Milestone's Achievements at Data Centers

500,000	300,000	250,000
HDDs Replaced	Assets Audited	Servers Relocated

Quite evidently, it's not viable and in most cases, not possible for every company to set up and run its own data center – given the prohibitive cost of real estate and challenges associated with the availability of adequate water and power. The good news is there are cost effective alternatives.

TRACK 2- Public Clouds:

Providing AI foundry and LLM capabilities through industry clouds in an open model

Businesses that lack the funds to set up their own AI data centers or colocation facilities but wish to build their own copilots and GenAI applications can do so using best-in-class development tools, frameworks, libraries, and LLMs provided by software platforms hosted on public and hyperscale clouds. Hyperscalers such as Microsoft, AWS, Google, Oracle, etc. offer the storage, compute, and cloud services, and the dynamic scalability, flexibility, performance, and cost optimization needed for supporting GenAI workloads.

Businesses can leverage NVIDIA AI Enterprise, a comprehensive cloud-native platform providing AI and data analytics software, along with an extensive library of frameworks and pretrained models. With NVIDIA AI Enterprise deployed on Azure, enterprises can draw from the computing power of NVIDIA's server and GPUs, and experience real-time enhanced GenAI inference predictions in models such as Microsoft Copilot. These cloud GPUs empower innovators to achieve the required hardware acceleration and gain access to massive computing power – on demand, remotely, and with ease. This is a growing trend, and according to IDC forecasts, by 2027, spending on accelerated AI servers in the cloud for inferencing is expected to exceed on-premises expenditures by more than threefold.

Businesses that seek greater customization can also connect their enterprise data to a suitable open-source foundation model and train it to meet their requirements. Today, there's access to prebuilt, out-of-the-box, customizable APIs and models for rapid innovation and development. These include the GenAI microservices offered by Azure OpenAI – that cater to the needs of developers and enterprises – offering Azure AI Document Intelligence, Azure AI Search, and Azure Machine Learning amongst others. In the health care industry, Azure OpenAI is helping patients quickly and efficiently connect with the right medical professionals at scale, while maintaining the quality of care.

Researchers can employ its generative models to examine medical images, identify irregularities, and develop new treatments. Its generative AI algorithms support disease diagnosis – helping analyze patient symptoms and medical histories and conduct precise and prompt diagnoses.

Microsoft and NVIDIA are two of several companies building their own AI foundries – with best-in-class models and implementation tools, and automation and industry capabilities built-in.

In the supply chain and logistics industry, GenAI is revolutionizing the predictive maintenance of fleets and equipment through the integration of IoT sensors. These devices collect vast amounts of real-time data from vehicles and equipment to analyze them and predict failures before they occur. The proactive approach minimizes downtime, extends the lifespan of assets, and optimizes repair schedules.

While this track doesn't exactly compare with the first one in terms of control and the extent of customization and flexibility, it nevertheless satisfies the need for high availability, performance, and data security. Businesses enjoy enterprise-grade support and services spanning software development, cloud computing, data analytics, data management, networking, and data security – without having to make deep investments.

Cloud service providers such as Microsoft and Google have also ventured to verticalize their cloud platforms – so businesses can benefit from specific solutions tailored for their industry. Referred to as industry clouds, these sector-specific services provide the distinct capabilities, tools, and services needed for businesses in the health care, banking, insurance, automotive, retail, and telecom space. Industry clouds offer the agility and ability for businesses to cater to their individual needs and accelerate outcomes.

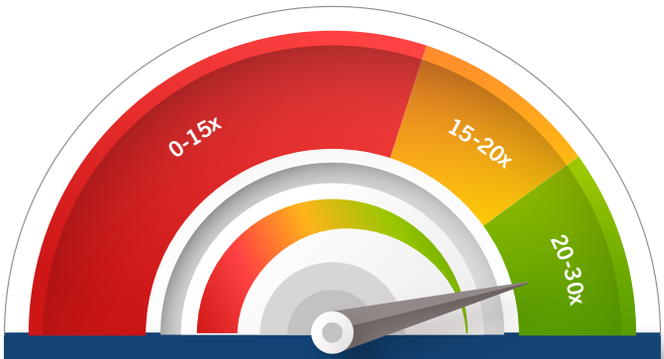
Emergence of Industry-Specific Clouds

Health Care | Banking | Insurance | Automotive | Retail | Telecom

Milestone conducts a thorough assessment of business requirements, identifies gaps and opportunities, and recommends the most optimal solution for businesses to create value from AI. Businesses rely on Milestone for making the transition to cloud environments and AI foundries. Milestone empowers businesses to manage multiple clouds – like it were one 'meta cloud'.

With Milestone, businesses are fully realizing the potential of Copilot in GitHub, Office 365, and other applications and systems. Milestone's project management, application integration and modernization, IT operations, and generative model training services have brought several benefits to businesses – such as a compute, configuration and cost advantages, rapid innovation, improved ROI, scalability, and competitiveness. Businesses have also earned greater trust of their customers, safeguarded their brand and improved resiliency, and developed the agility to respond immediately to new opportunities.

The Age of Meta Cloud



Milestone's Speed of Service Delivery

Milestone is rated as the best delivery experience and strategic partner by its clients and has delivered at 20 times to 30 times the speed of some global system integrators, resulting in significant and ongoing cost savings for businesses. Recently, Milestone assessed and remediated ~2000 AWS cloud services for an enterprise – significantly improving its operational reliability.

For businesses that are AI-reliant but can only go as far as subscribing to GenAI services and tools with plug-and-play and ready-to-go functionalities, there's the option of approaching SaaS companies that offer applications infused with GenAI capabilities.

TRACK 3- SaaS Platforms:

Digital Command Centers for enterprises of the future

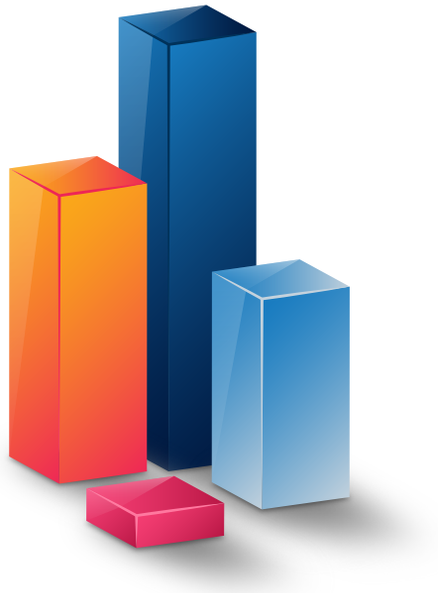
Businesses that aren't engaged in building their own LLMs or developing GenAI applications can study use cases pertinent to their nature of activities and approach SaaS players that fulfill their requirements through subscription-based models. SaaS players are bringing AI and workflow automation capabilities onto their platforms – enabling businesses to simply hit the ground running with GenAI applications within minutes. The convenience of getting 'GenAI in a box' through this pathway presents the lowest barriers to entry for companies that seek to swiftly embrace and possess GenAI capabilities.

SaaS-based GenAI tools and applications enable businesses to drive efficiency, productivity, growth, and competitive advantage in their respective markets. ServiceNow, for example, offers Now Assist – a GenAI powered service on its Now Platform that enables businesses to automate complex workflows, enhance decision-making, and streamline service management across various functions. This dramatically increases efficiency and reduces operational costs. Now Assist benefits diverse functions of a business – such as strategic portfolio management, IT operations, field service management, software development, HR service delivery, customer service, and IT service management. This creates capabilities such as auto-ticketing, auto-correlation, auto-escalation, auto-resolution, etc.

Along similar lines, the next generation Einstein platform offered by Salesforce – the leader in CRM solutions, uses ML and NLP to enable businesses to automate data entry, predict sales trends, and personalize marketing campaigns for improving the customer experience. Einstein empowers sales teams to prioritize leads that are more likely to convert and assists marketers to craft messages that resonate best with their audience – thereby driving agent productivity and increasing ROI. With Einstein Copilot Studio, businesses can also customize their Einstein Copilot to better suit their business needs. With Salesforce, businesses can launch new pricing promotions by studying the inventory in the warehouse and stores, ascertaining if any stock is nearing its expiry date, looking at what's perishable, and studying competitive activity. AI-enabled drones that circle around large warehouses and 'look' at what's going in and coming out can further assist businesses with their pricing and inventory mechanisms.

Businesses in the airline, hospitality, e-commerce, and ride sharing space are also turning to SaaS companies to deploy ML and train models for making real-time price adjustments – thereby accelerating profitability, improving market competitiveness, and saving precious hours for pricing practitioners.

By analyzing consumer behavior, supply and demand fluctuations, competitor pricing, and market conditions, GenAI models empower businesses with the capability to achieve dynamic pricing optimization.



Dynamic Pricing Optimization with GenAI

For creating marketing content aimed at specific customer segments or personas, businesses have the option of using AI assistants in HubSpot to generate fresh ideas and content for blogs, articles, websites, whitepapers, emails, social posts, etc. on a defined topic, besides improvising existing content. Concurrently, Salesforce can assist businesses with the new Configure, Price, Quotes (CPQs) and based on the SKUs, businesses can generate quotes rapidly and accurately for their orders. This convergence is a reality of today, where diverse systems, people and capabilities come together to create common business value.

To gain real-time actionable insights, conduct company research and keyword rankings, craft personalized responses, and generate reports on campaign performance, there's ChatSpot – HubSpot's conversational platform that comes in handy. By 2025, it's likely that 30% of outbound marketing communications from major organizations will be generated by AI – a significant increase from <2% in 2022.

Workday – a cloud-based company offering applications for HR and Finance, with its 65 million users, claims to have the world's largest and cleanest set of financial and HR data. Enterprises using Workday can tap into its GenAI capabilities to create job descriptions within minutes, analyze and create employee growth plans, generate personalized articles for employees, compare, and correct business contracts effortlessly, and much more.

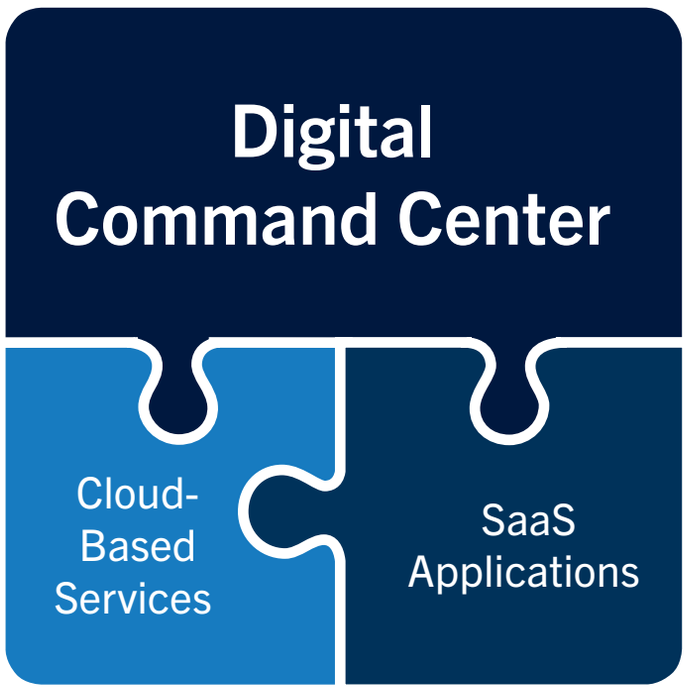
When it comes to AI-powered coding workflows, developers can turn to GitHub Copilot. By interacting with its chat feature and delegating routine coding tasks to AI, developers can focus more on innovation and problem-solving. With over 21 million US developers using GitHub to build applications, GitHub Copilot serves as a vital tool that accelerates their Software Development Life Cycle (SDLC), enhancing various parameters such as coding efficiency, security, and code quality. It assists with detecting vulnerabilities, maintaining consistency, and upholding software standards of an organization, thus improving developer satisfaction and overall project timelines.

Recently, NetSuite enhanced its GenAI offerings throughout its service portfolio, incorporating hundreds of new use cases across finance and accounting, supply chain management, sales and marketing, and customer support. The upgrade aims to boost the speed, accuracy, and efficiency of user operations. Through NetSuite Text Enhance, companies can utilize specific business data to generate and improve personalized and context-aware content. This advancement helps businesses achieve their objectives more swiftly and effectively by enhancing productivity, minimizing errors, ensuring consistency, and speeding up operational processes.

When looking for a solution to store all enterprise data in a common place and make it accessible to all concerned, Snowflake could be the answer. Think of it as an elastic data warehouse in the cloud with developer resources – where different teams can run simultaneous queries on the enterprise data and use it for analytics, application development, distribution, and scaling. Snowflake can be hosted on AWS, Google Cloud Platform, or Azure, and enables enterprises to use its container services to run leading LLMs and fine-tune them to suit their requirements. Using the GenAI capabilities of Snowflake Cortex, businesses can quickly analyze data and build their own applications.

GenAI is also revolutionizing IT Operations. With ServiceNow, IT Ops managers can receive pre-emptive failure predictions and other automated alerts neatly coupled with an analysis that pinpoints potential causes and recommends subsequent actions. This streamlined approach accelerates problem-solving with quicker and more efficient triage and resolutions, resulting in reduced Mean Time to Resolution (MTTR).

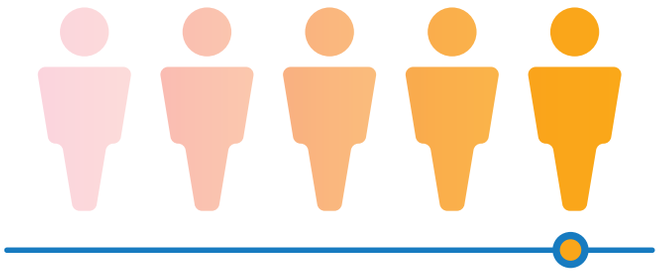
Enterprises of the future will realize that they can rapidly create a Digital Command Center (DCC) with almost no hardware investment – simply by deploying the right mix of cloud-based services and SaaS applications that meet their business requirements. Collectively, these will enable high availability of enterprise services, continuous observability and analytics, agility, and a comprehensive view of endogenous and exogenous events – enabling businesses to seize opportunities as they come and take remedial action on incidents that may potentially impact service availability. SaaS players such as ServiceNow will also serve as support models for data centers and multi-hybrid cloud centers.



Offered by third-parties, these applications and services need to be integrated with the business CRM, messaging platforms, databases, or other systems depending upon the solution being used. Some solutions represent low-code or no-code platforms but most of them need to be customized to maximize their potential and fully meet the strategic goals of a business.

For customization and deriving the most benefit, the technical proficiency and services of a strategic partner come into play. Milestone identifies business functions that have the greatest potential for performance increase through the adoption of GenAI. Through seamless integration, it enables businesses to maximize their GenAI capabilities, scalability, and ROI from software investments, while delivering consistent and high-quality customer experience across touchpoints.

Recently, Milestone delivered a Generative AI strategy and execution involving over 1000 users in less than three months. Milestone's rollout of GitHub Copilot in a large pharmaceutical company consisting of 1000+ developers led to a minimum 20% increase in developer output and 20% faster rate of code generation.



1000 Users in <3 Months

Moving onto our fourth and final track, we'll discuss how businesses engaged in building their own LLMs or training readily available models using APIs are leveraging GenAI for innovation and enhanced customer experience.

TRACK 4- The LLM ecosystem:

Should you build, buy, or use open source LLM? Efficiency/Productivity/Transformation use cases

Gartner believes that by 2025, more than 30% of new drugs and materials will be systematically discovered using generative AI techniques, up from zero today. Companies that capture, manage and use hypersensitive data – such as patient health records, financial transactions, investments, and others, and wish to leverage GenAI for creating business and customer value, will look at either building their own Large Language Model (LLM) or train an open-source LLM using APIs.

Pharma and life science companies tend to build and train their own LLMs for accelerated clinical trials, drug development, insight generation, and assistance in regulatory compliance. LLMs hold significant importance in data extraction, helping the research and scientific community to identify suitable patient populations. By working on historical data, the model can predict not just drug interactions, but also the studies and clinical trials that are most likely to succeed. Accordingly, companies are making necessary design adjustments in their trials to maximize the likelihood of success while avoiding costly failures. By creating digital twins of patients using GenAI, researchers are simulating control group scenarios. This significantly reduces the number of actual control patients needed in clinical trials – potentially to just half.

Wealth management companies, financial services firms, and fintech companies are also increasingly utilizing LLMs to develop robo-advisors and recommendation engines to enhance their services and improve operational efficiency. Working on the large datasets of these businesses, GenAI can analyze historical data and predict short and near-term price movements. It can provide information for trading in real-time with indicators for making buy or sell decisions. Integrating LLMs with open-source technologies opens the doors to a world of possibilities – such as automated financial analysis, enhanced real-time decision-making in trading, deeper insights into sentiment analysis, and highly personalized investment strategies for clients.

The decision of “build vs. buy vs. open source” in LLMs depends on several factors, and each option has its own benefits and challenges. Buying a pre-trained LLM model offers convenience and saves considerable time and cost in building and training. However, it offers the least freedom and flexibility for further customization. A pre-trained LLM will need fine-tuning – to align it to the industry standard and customize it for specific business needs, and this requires specialized skills.

On the other hand, open-source LLMs offered by the likes of OpenAI, Mistral AI, and Anthropic can be trained through LLM APIs, and offer greater control and room for customization, but may not completely align with the business needs, and require special skills and greater time for training.

Lastly, building an LLM, while being the most time and cost intensive option, offers maximum customization and control over the dataset. Effective AI implementation requires training data, foundational data sets, and robust infrastructure, including specialized chips, processors, memory, storage, and GPUs. Just for context, GPT-3 was trained on around 45 terabytes of text data.

Businesses would need to ascertain the size of the training dataset, compute, memory and latency requirements for developing and running their GenAI applications, and accordingly make investment decisions. Instruction tuning may be necessary in most cases, where the LLM is further trained on a special dataset for following instructions correctly. Fine tuning can also be carried out through prompt learning to further refine the model by guiding it to learn specific tasks using P-tuning and prompt tuning methods.

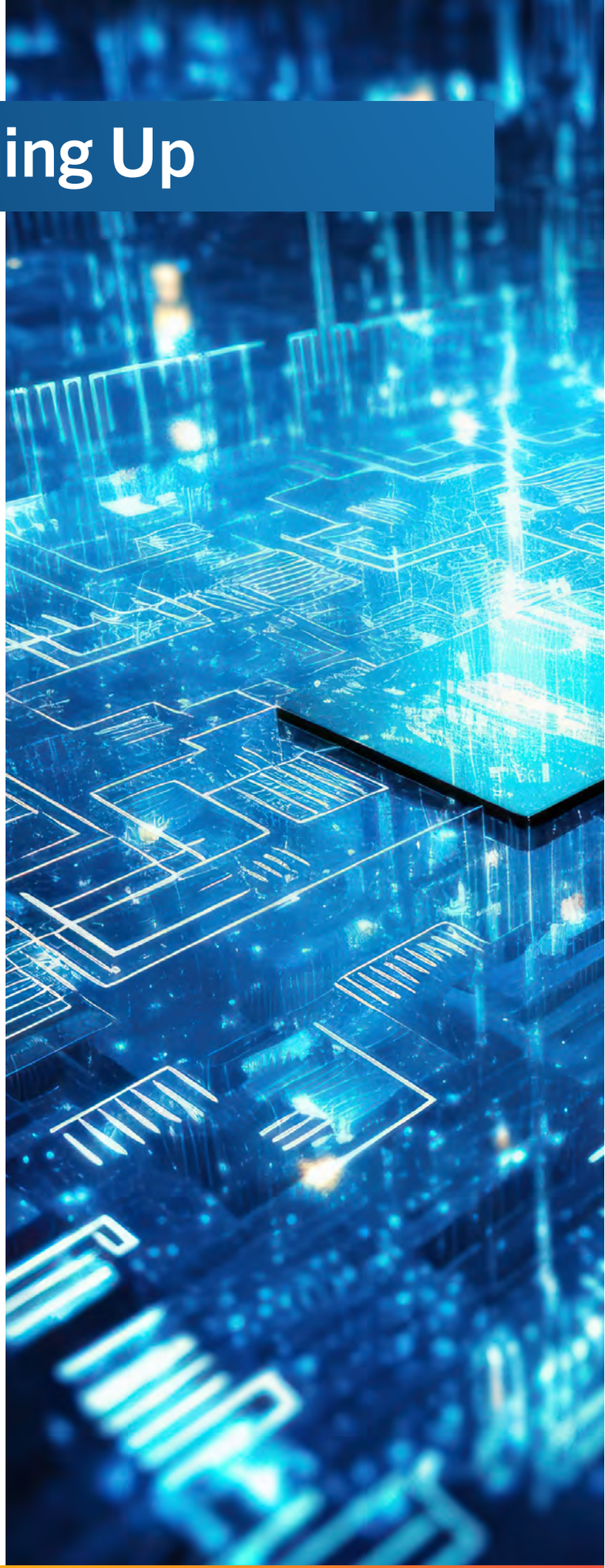
With over 25 years of experience working closely with emerging technologies, Milestone has deep expertise and sound capabilities in AI. Milestone guides businesses towards building LLMs, training them, and generating the most value for themselves and their customers.

Summing Up

The diverse applications of GenAI across business functions underscore its transformative power, enhancing operational efficiency, accelerating innovation, and providing competitive advantage. With GenAI, businesses can exponentially increase productivity by creating tailored content almost 10x faster – from writing and copy editing to advertising and compiling meeting notes.

Innovation thrives as GenAI aids in new drug discoveries, optimizes clinical trials, and powers digital solutions in FinTech and healthcare. Collaborating with a strategic and implementation partner enables companies to fully capture GenAI's value, paving the way for greater achievements and sustainable growth.

Wondering where to start with GenAI in your business? Or how you can exploit the full potential of your technology investment? Connect with us and let's discuss how we can move the needle further on innovation, agility, cost and process optimization, productivity and other parameters that matter the most.



About Milestone

Milestone Technologies is a leading global IT services and digital solutions provider that collaborates with organizations worldwide to revolutionize their technology infrastructure and digital capabilities.

The company specializes in providing solutions across Application Services and Digital Engineering, Digital Workplace, Cloud, and Infrastructure Services, AI/Automation, Business Process Services, Salesforce, and ServiceNow.

With a strong commitment to innovation and customer satisfaction, we empower businesses to accelerate their digital transformation journey and unlock new opportunities for growth and success.

By leveraging our extensive expertise in cutting-edge technologies, we provide companies with the agility and scalability needed to stay ahead in today's rapidly evolving digital landscape.

Our comprehensive suite of IT services encompasses everything from cloud solutions, AI, and application services and consulting, to managed services, cybersecurity, platform engineering, data analytics, digital workplace services, and ServiceNow, enabling organizations to optimize their operations, increase efficiency, and drive value.