White Paper

Organizational Transformation Leveraging Modern Infrastructure to Deliver Cognitive, AI, and Analytics Capabilities

Sponsored by: IBM

David Schubmehl Larry Carvalho Donna Nitchie

April 2018

IDC OPINION

The Power of Cognitive Applications

Gaining insights from rapidly expanding volumes of data residing in a variety of sources is challenging enterprises to meet pressing organizational demands, which often include speed and agility. Traditional approaches to deep learning and cognitive computing were complex and reserved for scientific and very large organizations with equally large budgets. Yet advancing innovations in technology are enabling applications to make increasingly complex and contextual decisions that are timely, accurate, and precise. At the foundation of this are modern cloud infrastructures built on accelerated compute processing technologies such as GPUs. The evolution and proliferation of cloud technology have simplified and made more economical delivery of these IT services possible as vendors combine infrastructure and software to enable consumption of cognitive capabilities to far more organizations than ever before.

Organizations within many diverse industries now have the potential to leverage cognitive applications and transform the way they process information to interact with their users. Large volumes of unstructured information are generated every day in the legal and medical fields for instance, and providing rapid insights from this information transforms its value and opens the potential for new revenue streams and business outcomes that promise greater customer satisfaction. Delivering cognitive applications that offer these capabilities as cloud-delivered services reduces the cost and staffing expertise barriers to making these types of capabilities a reality.

The ability to process information immediately enables organizations to make real-time decisions and react to the fast-changing business environments. The speed at which new infrastructures can deliver insights based on a stream of new information will impact many use cases such as developing dynamic advertising campaigns and addressing widespread and complex security challenges. Numerous business processes can be automated based on real-time information. In many cases, the insights gained via cloud-delivered cognitive applications are simply not possible via a more manual application because of the sheer volume and velocity of data involved.

Embedding Cognitive in Existing Applications

Enterprises have built applications over the past few decades that use a variety of technologies and frameworks that have improved over time. Maintaining these applications while simultaneously

embedding new functionality or creating new applications that coexist with the old is a constant battle for application development managers. API connectivity is critical for enabling the building of cognitive capabilities for these existing applications. New delivery and consumption of cloud-delivered deep learning services open new methods to automate business processes and give organizations the ability to meet line-of-business demand.

A Rapidly Growing Market

IDC's digital transformation research predicts that by 2020, 85% of new operations-based technical position hires will be screened for analytical and artificial intelligence (AI) skills, enabling the development of data-centric digital transformation projects without hiring new data-centric talent.

IDC believes this rapid adoption will continue to be turbocharged in 2018 and beyond by the emergence of deep learning and AI-as-a-service offerings that will be more widely available in the marketplace.

Developers are the critical user population that will impact the speed at which AI will take root in enterprises over the next several years. IDC predicts that cognitive computing, artificial intelligence, and machine learning will become the fastest-growing disciplines within software development by 2019 and that 90% of enterprise development teams will be using cognitive/AI and machine learning tools and services as part of their toolsets by 2021.

Enterprise use cases that will have the greatest traction over the next 36 months are as follows:

- AI services embedded in (or added to) existing applications
- Advertising/marketing (hyperpersonalized customer experiences)
- Manufacturing and retail supply chains
- Design/engineering
- Customer support
- Asset management

Several of these use cases will be within the IT organization itself. IDC predicts that by 2021, 50% of enterprise infrastructure will employ some form of cognitive and artificial intelligence to improve enterprise productivity, manage risks, and drive overall cost reduction. Cybersecurity will also be a critical first-wave application of AI. IDC predicts that by 2020, 60% of the G2000 will use AI-based security.

SITUATION OVERVIEW

The market for AI and deep learning applications is surging. IDC estimates that spending on AI and deep learning solutions will exceed $57 billion by 2021 and that 75% of all enterprise software will include some aspect of machine/deep learning for predictions, recommendations, or advice by 2026.

In thinking of these applications now or in the near future, organizations should consider the following best practices and helpful guidelines:

- **Encapsulate and "systematize" best practices.** This is a variation on the themes about learning from experience. Developing machine learning models that replace rule- or heuristics-based

systems is a key use case in this area. Establishing metrics and measures that can be applied across models and training data will drive accuracy of the model.

- **Personalize outcomes and recommendations.** Many organizations are beginning to use deep learning models to "personalize" content, predictions, and recommendations to specific customers or prospects. This is especially true with mobile applications, where users increasingly expect their devices and applications to "know" their likes, dislikes, and expectations.
- **Prioritize security.** It is critical that organizations be prepared to manage the data about every user across every channel used to gather their data. Concerns about privacy and regulation of sensitive and personally identifiable information continue to increase and drive regulation (e.g., the EU's General Data Protection Regulation [GDPR] and the mining of Facebook data by Cambridge Analytica in the United States.)
- **Develop a corporate data strategy for AI and machine learning.** To personalize outcomes and recommendations, the deep learning models mentioned previously are based on data, lots of data. Most organizations are sitting on piles of data that has been collected over the years, from customers, suppliers, competitors, and their own business analysts. This data can be reused as a source for machine learning and predictive applications. In addition, third-party data sources can augment and enhance the available first-party data. Organizations need to develop a corporatewide strategy for the full life cycle, birth to death, of this wealth of data. Many enterprises have created C-suite roles for data governance, such as a chief data officer (CDO), or expanded the role of the CIO.
- **Augment human judgment.** The best business cases are about extending human capabilities, not replacing them by positioning AI-enabled applications as an extension of human intention. Power tools in the hands of a craftsperson is a good analogy. Pricing optimization models are examples of deep learning in this area.
- **Accelerate investigation and discovery.** Even the very best human readers can't ingest millions of pages of documents in one day. Applications that understand natural language can also be applied to this task for the spoken, rather than just the printed, word. Deep learning–based natural language tools and systems (e.g., natural language processing [NLP], sentiment analysis, named entity recognition, and speech-to-text [STT]) provide better results than handcrafted taxonomy-based systems.
- **Recommend "next best actions" and predict outcomes.** Deep learning–based applications build models using relevant data for recommendations and predictions, which are some of the typical use cases. For example, drug interaction models can help pharmaceutical companies and weather models can help retailers explain why some products sell better at certain times than others. Seemingly unrelated data can affect consumer behavior in unexpected ways, and developing comprehensive models is one way to reduce the uncertainty.
- **Automate organizational knowledge management.** While knowledge management systems have existed for decades, many have failed under the weight of human effort required for ongoing operation. Applying automation to investigation and discovery activities or developing best practices is a key benefit. Subject matter experts (SMEs) are used in effectively training automated systems, bridging human-to-machine intelligence. Automatic categorization and theme identification are some of the key use cases of deep learning.

Organizations are using deep learning applications as a catalyst for business process disruption, digital transformation, and the creation of new economies of scale. Large healthcare organizations are examining how deep learning applications can help democratize and accelerate "best practice" diagnosis and treatment regimens, no matter where their clients live. Global financial institutions are

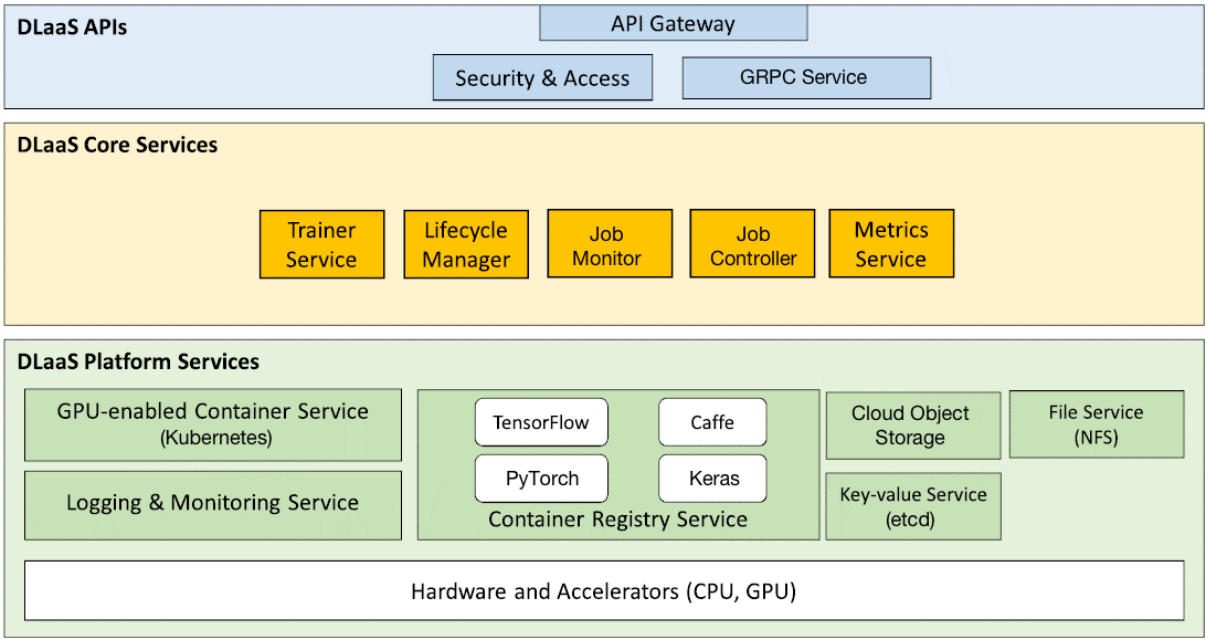
using AI-enabled applications to accelerate, automate, and sometimes eliminate manual workflows and business processes that handle financial transactions. Manufacturing companies are developing sophisticated predictive maintenance strategies based on IoT and deep learning models as well as revolutionizing just-in-time manufacturing with dynamic sourcing of raw materials.

DEEP LEARNING IN WATSON MACHINE LEARNING

IBM, long a leader in cognitive technology, has brought its capabilities to link big data with powerful hardware accelerators (e.g., NVIDIA GPUs) in the flexibility of a scalable cloud infrastructure, with the addition of deep learning with Watson Studio as part of Watson Machine Learning. With the new deep learning service, IBM shows how cognitive computing no longer depends on high-performance computing (HPC) environments, making it accessible to and affordable for many organizations. Available as a service from the IBM Cloud catalog, Watson Machine Learning enables data scientists in building new or enhancing existing applications with deep learning capabilities, as well as application developers who are typical consumers of deep learning models. The architecture of the system is shown in Figure 1. Designed with microservices for rapid deployment, the new deep learning ability within Watson Machine Learning simplifies access to the underlying hardware and software infrastructure.

FIGURE 1

Deep Learning Service Architecture



Source: IBM, 2018

Platform Layer

The platform layer manages both the deep learning components and the machine learning model training jobs. Functions include configuring the platform environment, linking to datasets and deep learning frameworks (currently these include TensorFlow, Keras, PyTorch, and Caffe), managing container services and credentials (e.g., Docker Registry and Kubernetes resource manager enhanced for GPUs and Mesos/Marathon), and executing and monitoring the training process. This layer also handles software provisioning and updates (much software is open source) while managing the infrastructure for scaling and resiliency.

The Watson Machine Learning microservice-powered platform enables the deep learning processes to be distributed and coordinated across GPUs and CPUs and across machines in a dynamically scalable heterogeneous cloud environment.

API Layer

Via REST APIs or the command-line interface, the system interacts with users or other applications. An API service registry enables load balancing and provides failure management; failed nodes are reallocated and failed jobs are retried.

Core Services

At the heart of the deep machine learning are IBM's core microservices for deploying, training, managing, and measuring jobs from submission to completion. These core services meet the unique needs of deep learning cloud-based cognitive analytics. Unlike typical cloud applications (such as transactional applications or web searches) that are short running or stateless, deep learning jobs require complex parallel computation and can run for long periods of time. This requires services that provide resiliency from expected failures in infrastructure (e.g., network congestion and planned upgrades/outages) and data persistence in a distributed compute environment.

Experiment Assistant

IBM Studio provides visualization to view and track model training progress such as accuracy and loss measures. This enables users to see trends, patterns, and anomalies at a glance and make decisions on adjusting the models in a timely manner. Metrics are collected, parsed, correlated across logs (e.g., trainer and GPU utilization logs), visualized, and rendered in real time. Microservices such as Rickshaw allows for creating interactive time series graphs.

IBM is developing Knowledge Studio and IBM Data Refinery to include additional data capabilities around data ingestion, cleaning, and inferencing in Watson Machine Learning. The company is also extending the visualization around the training behavior to increase user interaction in the process. The deep learning infrastructure made available to customers relies on the same underlying services that power IBM's other cognitive services like Watson Visual Recognition and Watson Natural Language Classifier.

IBM's Watson Machine Learning merges deep learning and abstracted infrastructure and so brings a rich level of cognitive capabilities to organizations in the process of digital transformation. With capabilities which previously required high performance and even supercomputing environments now available via the cloud, Watson Machine Learning puts cognitive deep learning capabilities well within the reach of diverse applications and industries. The economical approach of cloud-based services provides the ability for organizations of almost any size to experiment and develop solutions that are cost effective and don't require capital expenditures or specialized AI skills.

CHALLENGES/OPPORTUNITIES

The biggest challenge facing IBM and other AI platform vendors is that almost all cloud platforms are adding AI and deep learning capabilities to their capabilities as the cloud applications of the future will all be what IDC calls "AI enabled." That is, these cloud applications will incorporate aspects of machine and deep learning to provide recommendations, predictions, and prescriptive advice based on models that have been built with data rather than using rule-based or heuristic algorithms.

As all of the major cloud platforms incorporate AI and deep learning, IBM will need to contend with this very crowded marketplace and offer services and capabilities that are superior in meaningful ways as well as offering superior AI and deep learning capabilities. For IBM, the opportunity is immense as organizations transition to the second generation of the 3rd Platform, where hyperpersonalization and customization based on AI become table stakes in the next round of consumer and enterprise applications that will emerge over the next two to three years.

There is a scarcity of developer skills in using and deploying AI technologies in applications to enhance organizational capabilities. While IBM has all the components that make it easier to adopt the technology, the company needs to convey this message to developers by providing an educational path to helping implement common use cases.

CONCLUSION

For enterprises, AI-enabled applications represent a methodology to perform business processes better, faster, and more reliably than ever. AI-enabled applications will provide disruptions to many traditional enterprise business models, especially in healthcare and ecommerce. Enterprises need to be aware of this and start thinking about how they are going to deal with future disruptions that come about because of cloud-based AI technologies and capabilities.

Recommendations for organizations considering AI and deep learning as a service vary depending on where they are on the journey. IDC advises creating an AI center of excellence to encourage discovery, learning, and cross-organizational collaboration between businesses, IT, and data scientists. Plan for creating technology capabilities — including platforms, technologies, processes, governance, talent, and data components — that will empower the enterprise. CIOs must create and continuously enhance an integrated enterprise digital platform that will enable new operating and monetization models. Enterprises should actively consider and plan for AI-enabled applications within their organizations and/or develop plans for consumer-facing AI-enabled applications. Organizations should consider developing prototypes and pilots using cloud-based AI technologies such as those found in the IBM Cloud. IT will need to ensure the availability of best-in-class API management tools that interface with data feeds for AI-powered applications and services. APIs are critical to machine learning and cognitive apps as a means of overlaying multiple data sources to enhance analytic accuracy and the sophistication of data-driven insights.

AI-enabled applications are fueled by data. Organizations should undertake processes and plans that identify key data sets and repositories that will be required for AI-enabled applications to be successful. This may also require looking externally at third-party data such as that offered by IBM's Weather Company or offered through Watson Health. Privacy and security will present serious land mines for AI-based digital transformation efforts. Define robust privacy and security rules for AI-enabled applications that dictate what data can be collected by these applications and the set of permissible uses for that data. Stay abreast of the constantly changing global regulatory environment.

Organizations that can adapt quickly to regulatory changes will be at a significant competitive advantage to others in the market that don't adapt as quickly.

AI-based applications, processes, and services are changing the world. Organizations need to have a strong AI strategy that makes the best use of resources while protecting and securing the enterprise from threats and malefactors. Deep learning-as-a-service offerings such as IBM Cloud provide a safe and secure method of achieving this while providing agile, economical tools to build and run AI-based applications.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com