White Paper

AI IN CLOUD 3.0 AND ROADMAP FOR ADOPTION

ANDREI G. STOICA, PH.D., Senior Vice President Software Development, IQVIA

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Cloud 3.0 marks the beginning of Artificial Intelligence (AI) cloud orchestration. Commercial-grade AI requires AI algorithms, domain subject matter expertise for modeling, big data for training and massive compute on specialized hardware. The ecosystem of web services and serverless compute in Cloud 3.0 is accelerating the hyper-convergence of AI building blocks. To capitalize on these technologies, CIOs and CTOs must reengineer the process to build Machine Learning (ML) and AI applications. With a few changes in application design patterns, organizations can begin to access previously unreachable expertise, data and compute to build disruptive applications.

AI AND ML BASICS

Although the following is not an in-depth description of AI, it presents the basic concepts to start the AI journey:

- **AI** creates machines that mimic cognitive human functions.
- **ML** is a subset of AI that builds a model using training data and then applies the model to new data. The goal is to create a predefined output such as decisions, predictions, classification or other analytics. An example of AI that is *not* ML is a rules-based engine.
- **Neural Networks** are a class of ML that mimic the behavior of biological neurons. The input is a digital model of the domain knowledge and the output is a digital model of the result, such as a classification, decision or prediction, etc.
- **Deep Learning** is applied to neural networks to learn abstract concepts progressively. At each layer, the neural network learns new concepts (such as the meaning of a sentence) using aggregated concepts from the previous layer (such as the meanings of single words).
- **Supervised Learning** is used in ML to teach the system how to behave when we know the results for a set of training data. The system learns to mimic human behavior from a given set of inputs and outputs.
- **Unsupervised Learning** is used in ML to learn the structure of a data set when we don't know the results even for the training data. Unsupervised learning is used in clustering, for example.
- **Natural Language Processing (NLP)** is the application of multiple ML models and training methodologies to understand natural language.
- **Graphical Processing Units (GPU)** are hardware components that are optimized for processing images using massively parallel computation. While the Central Processing Unit (CPU) in a server has up to tens of cores (processors), the latest GPUs have up to thousands of cores.

ROADMAP TO ADOPTING AI IN CLOUD 3.0

The basic concepts of AI can be overwhelming for IT professionals who don't have years of data science education and experience. However, with Cloud 3.0 it can be relatively simple to adopt AI quickly. The first step is to **re-engineer the enterprise architecture and software development process** to Cloud 3.0 and to **maximize the use of web services** (see "Evolution to Cloud 3.0 and Roadmap for Adoption" and AccessPoint, Volume 7, Issue 14 for practical steps to adopt Cloud 3.0). This will change the fundamentals of application development and provide novel ways to connect the AI building blocks.

The next step is making the **build/buy decision for the AI building blocks.** AI algorithms are widely available in commercial or open source tools and libraries. Compute capacity on specialized GPU hardware is also becoming easily accessible to rent or acquire. Data is typically a combination of purchased industry data and enterprise data.

Finally, **addressing domain subject matter expertise** is the most difficult of the AI building blocks. Domain subject matter experts know the enterprise business, how it is reflected in the data, and how to harness it with AI algorithms. Organizations may build multidisciplinary AI teams as well as buy AI web services. Multidisciplinary teams are composed of data scientists with in-depth AI science expertise and some domain knowledge who work alongside experts with in-depth domain knowledge and some AI science expertise. Such internal teams are necessary but, in many cases, they cannot cover all of the organization's needs. This is where Cloud 3.0 disrupts the current AI application development. By designing AI apps as a series of interconnected microservices, application architects can take advantage of the emerging Cloud 3.0 marketplace of web services.

CIOs and CTOs must create an **AI-specific design track for the microservices enterprise architecture** process. A typical app in Cloud 3.0 is a collection of microservices, built in-house or commercially sourced. For example, designers have the flexibility to purchase a training web service and ML models from a big data provider and build in-house microservices to do the inference on the enterprise's own data. Using a commercial web service for training can be a much faster and more cost-effective approach than purchasing the data and building the training in-house.

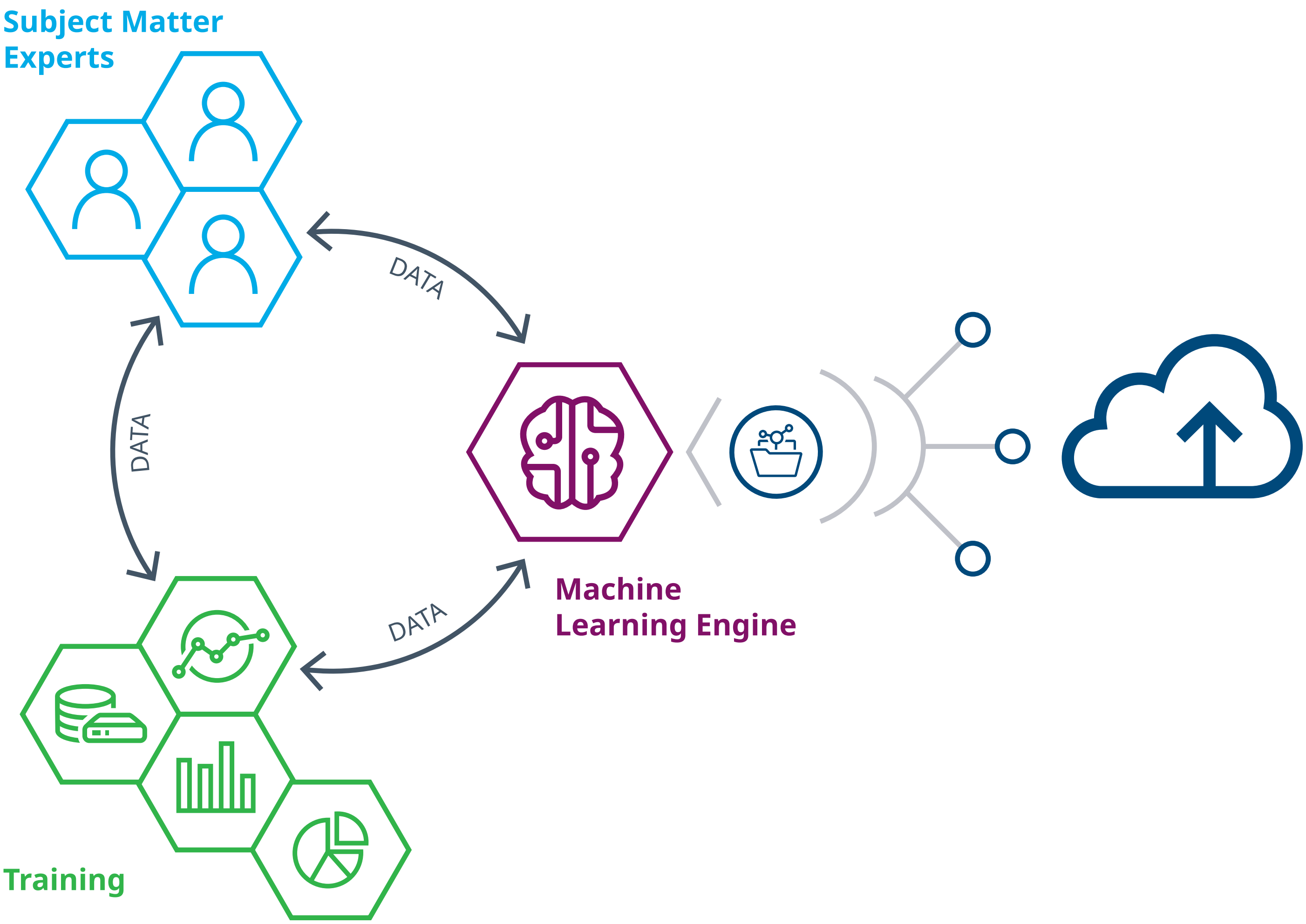
Below are the major constraints and dependencies for microservices design and vendor selection:

- **Training vs Inference.** Training requires massive amounts of computing power and data to build the ML model. The model is then applied to new data for inference in support of a business process. Inference requires less storage and compute, by orders of magnitude. The apps required for training and inference can be separated.
- **Location, Location, Location: On-Prem, Fog, Cloud.** On-Prem (also known as network, core or ground) is the enterprise data center. On-Prem computing is considered when servers on the internal network host the enterprise data and processing requires heavy data exchanges between multiple apps. Fog computing (also known as edge computing) is the distributed computation that happens on remote servers or devices. Fog computing is considered when the algorithm allows for distributed pre-processing of data originating from the edge nodes. Cloud computing is considered when there is light data exchange between the apps or data is already in the cloud. The major design considerations include bandwidth and latency requirements, service level agreements (SLA), the total cost of ownership (TCO) and the speed to market.

- **AI Algorithm Design.** Some AI algorithms allow computation to be distributed to a large number of nodes where each node is working on a subset of data. This breaks the data in small subsets – the computation is in parallel, and results are aggregated at the end. Faster computation can be achieved by adding more processors and GPUs are extensively used in these cases. Other AI algorithms require a single high-power computing node processing the entire data set. The algorithm dictates the bandwidth and latency for communication (which in turn determines location) as well as the size and number of the computing nodes. Ensemble methods use multiple AI algorithms on the same problem, allowing increased flexibility in the app design.

MEN VS. MACHINE ON THE AI JOURNEY

One of the common fallacies of the AI journey is that a “killer app” can replace the human operator. While this is a desirable long-term goal, the first practical milestones must be automation and introducing AI as an aid rather than a complete solution for the target business process. Practitioners will need to address the **challenges of change management and organizational inertia** in addition to the effort to develop advanced AI applications.



Successful AI systems rely on humans for two distinct functions. First, domain subject matter experts must constantly **refine the AI algorithms and tune the training process.** The input data for training, the algorithms and the timing for training are essential to the performance of the AI app. Second, humans must constantly evaluate and validate the output (see “How Machines Learn in Healthcare” in AccessPoint, Volume 7, Issue 14). The more quality training data that is validated by human operators and fed back into the systems, the better the AI will perform. The art of building AI apps is in the design of the data flow for training and validation as well as in the balance between humans and machines. Practitioners should start with recommender systems that learn in the background. Upon confirmed performance, the app architecture should evolve towards more and more autonomous AI operations.

CONCLUSION

Because of its intrinsic complexity, the AI journey is fraught with misconceptions, unreasonable expectations and inflated marketing claims from vendors. At the same time, it holds the realistic promise of disrupting the marketplace faster and more substantially than past technology waves. Organizations can build low-risk AI adoption roadmaps using Cloud 3.0 design patterns by starting with automation and slowly adopting more autonomous AI. To fuel innovation for the organization, successful AI systems must have a balanced man-machine approach that is incrementally evolving.

ABOUT THE AUTHOR

ANDREI G. STOICA, PH.D.
Senior Vice President Software Development, IQVIA

Dr. Andrei Stoica is responsible for designing and implementing the data cloud platforms that support the delivery of the company's information offerings. Under his leadership the company has developed one of the top three high performance private big data clouds in the world, hosting the largest and most diverse healthcare sets. Dr. Stoica has been with IQVIA since 2006 holding several leadership positions in information technology, software development, technology innovation and information security. Prior to joining the company, he served in information technology roles in the healthcare IT industry, as well as research and teaching positions in academia.

CONTACT US

iqvia.com/contactus

IMS Health & Quintiles are now IQVIA™

Copyright © 2019 IQVIA. All rights reserved. WP.0098-1-07.2019