AI Performance Management:

Engineering AI applications at the enterprise level

WHITE PAPER

Introduction

As the world is moving towards ubiquitous connectivity, everything -- from devices, machines, cameras, and humans – is generating humongous data points like logs, audios, images, videos etc.

Organizations are now analyzing these data points to extract intelligence and create a new range of services. Building intelligence involves analyzing these data sets for patterns with the help of technologies like Artificial Intelligence, Machine Learning, and Deep Learning to create factories of future, autonomous vehicles, smart/safe cities, smart farming, etc.

Typically, the process involves AI engineering team to define the problem, create the application architecture, choose the infrastructure, data sets, AI framework, and the build trained model, deploy it to the right node (web/mobile/machine), explain the model to business users in a non-technical way and setup the governance to repeat this process. New data sets are used to train models continuously and the models are monitored for consistency and accuracy. Organizations struggle to build the right AI solution as it involves the complex task of choosing Infrastructure, data sets, AI frameworks, monitoring tools, governance procedures, explainability tools, and more importantly the skilled resources.

Organizations working on large-scale AI Initiatives are hitting roadblocks with their timelines and cost constraints due to multi-dimensional complexities involved in handling/developing AI applications. An AI engineer working on autonomous vehicles has to analyze about 10bn miles of video data. In the case of drones, each drone flight generates about 450 GB of data that moves from storage on the drone to either cloud or datacenter and the data is used to train AI models to inference meaningful insights. Intelligent surveillance cameras have capabilities to do the inference on the edge while also transferring the data to the data center/cloud as well as import the trained models back to the camera, frequently.

Organizations struggle to build good AI solutions due to multiple moving parts in the process and discover complexities during development lifecycle to choose components, subsystems, frameworks, tools, accelerators, and diverse skillsets. We discuss how to approach building an enterprise-grade AI application within the given constraints.

Overview

Worldwide revenues for AI are expected to be 300 Bn by 2024 and are growing at 17.1% CAGR according to IDC.[1] Typically, productionizing the trained model and performance tuning is half the cost of any AI project and organizations need expensive skills to get meaningful outcomes for AI Investments.

AI solution requires the selection of a right dataset, infrastructure, algorithms, and frameworks to train and build the trained model. The Hardware platforms are continuously evolving with options to choose CPU/GPU or other accelerators and abstract the hardware from the developers with right libraries and standards. Similarly, there are new algorithms and evolving frameworks available to developers adding complexity to choose the right fit. Finally, the knowledge required to select these complex build blocks is distributed and requires good coordination to setup the right process of build a reasonably good-trained model.

The trained model is packaged so that the end-user can use it to inference on different environments and meet the business objectives. In some cases, while using these trained

[1] https://www.idc.com/getdoc.jsp?containerId=prUS46757920

models for inference, users will bring in new data that will be fed back to train the algorithms and improve the model accuracy and reliability. The governance around the entire development process is key for shorter development timelines, as good processes will allow successful repetition of experiments and tuning the model parameters.

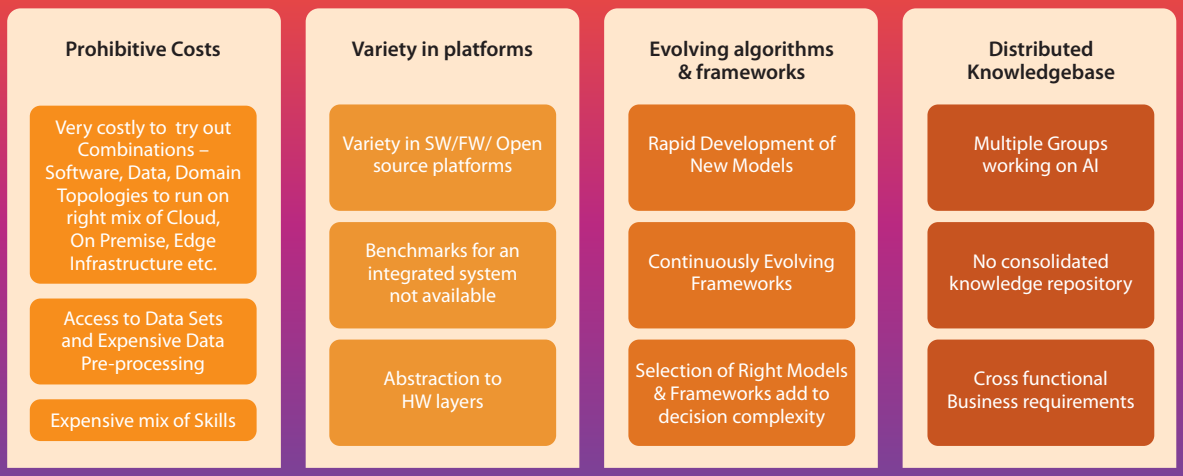

Figure 1: Barriers to Robust AI Application Development

AI Performance Management

Highlighted below are some of the key issues that AI stakeholders need to consider for accelerating their AI Programs.

Datasets

Finding the right datasets is often the biggest challenge for an organization intending to build the AI solution; open data sets are only somewhat useful. For example, a team working on disease prediction by observing eye patterns can depend on an open data set but the model built with such data may be accurate to the geography in which data is collected. It is important to gather right datasets that are free from privacy concerns, meet confidentiality, and regulatory requirements (like GDPR) and constructed with necessary consent from users to build applications. It is also important at this stage to establish the data governance and ownership around the data sets.

After compliance requirements, the choice of data store becomes the next key consideration. Given dataset volumes, it is often expensive to store, retrieve, and move data while simultaneously massaging the data for consistency and managing the data pipelines.

Infrastructure & Network

The choice of CPU, GPU, FPGA/ASIC, network, storage, and bandwidth are key components of infrastructure and the choice depends on the type and size of the problem being solved.

For instance, noise removal in an audio sample would require a selection of CPU-heavy infrastructure, while an image-processing problem needs a GPU heavy infrastructure. The algorithms can also be ported on hardware accelerators like ASIC for better processing speeds but with some loss of flexibility to change it continuously.

Data movement and size will determine the options for network, storage and bandwidth and will heavily impact the training process as often large datasets are moved across devices during the data preparation phase.

The proportion of CPU/GPU/DSP/FPGA/ASIC, and a good selection of network, storage, bandwidth, and more importantly the right mix of skills to manage infrastructure is critical to achieve optimization of training time, model performance, power efficiency and accuracy.

Cloud service providers offer users to select the multiple combinations to bring up the infrastructure required for application development.

Algorithms, Frameworks

Video analytics, language translations, noise removal, and NLP offer a variety of AI development challenges and there are numerous algorithms and frameworks to help the AI development process. Machine learning algorithms and neural nets are evolving and adding complexity to baseline the development of AI models. AI development involves mastering the application development in one or more frameworks and selecting the right algorithms. Developers need to iterate and experiment multiple times with tuning parameters to achieve accuracy, engineer to compress the models to fit on the target platforms, and monitor algorithms efficiency with new data sets.

It is also important to choose the interoperable standards for models so that switching with frameworks does not consume huge resources or skills. Sticking to a popular standard like ONNX (open neural network exchange) will help users to easily port the models between frameworks. Python is generally the choice, but users can choose c/c++ where execution and response times are critical.

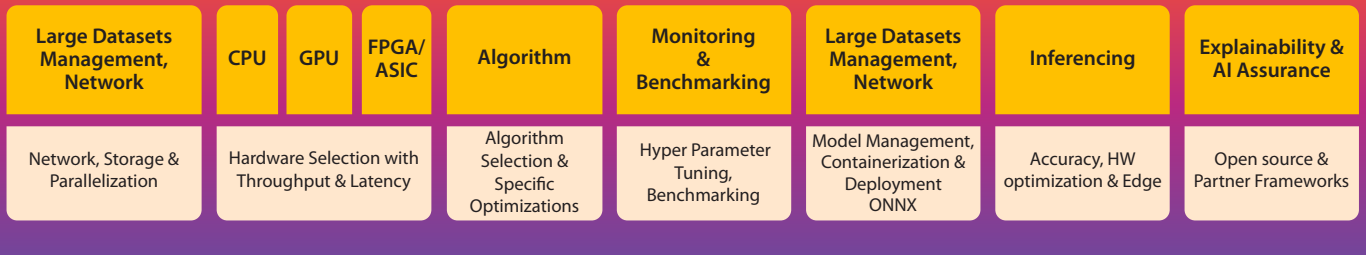

Figure 2: AI engineering building blocks

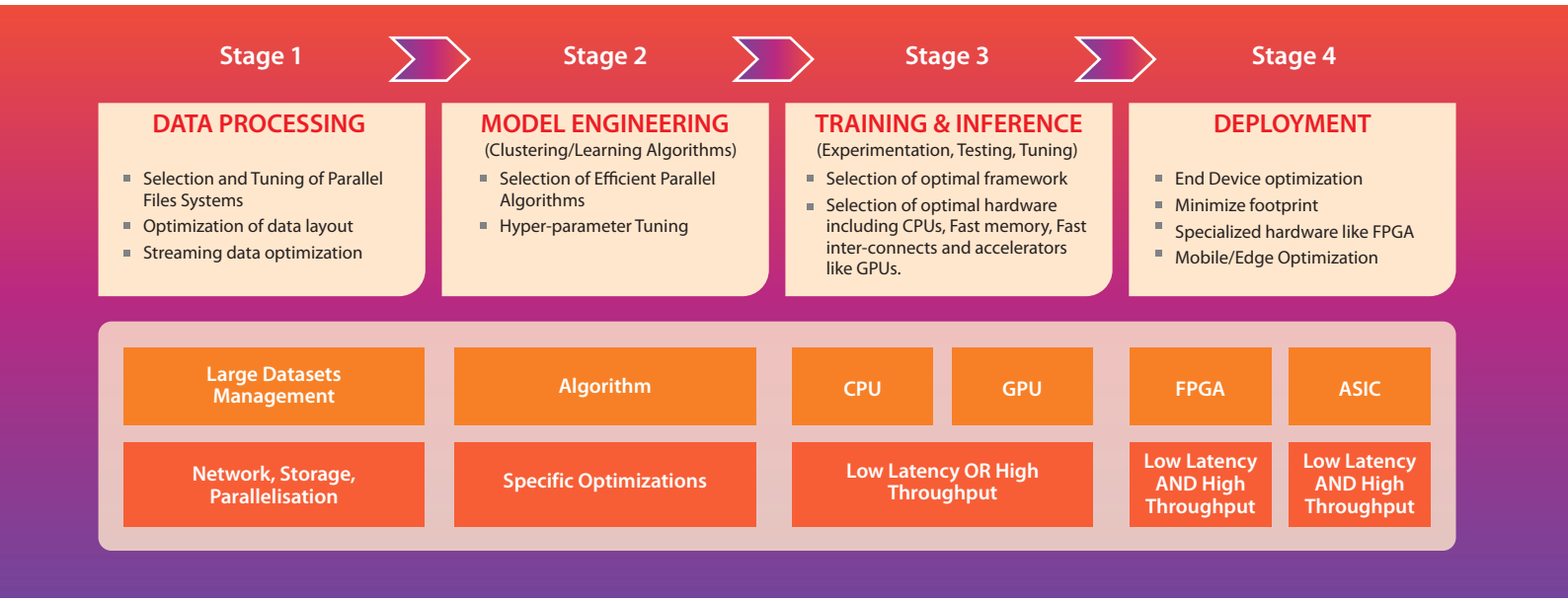

Figure 3: Acceleration at different stages

AI deployment, model management & governance:

AI application development, infrastructure setup, data governance, training, model generation, model management requires consistent governance practices for repeated experimentation. Even the new datasets keep adding to the original data sets during the AI application development process. AI developers perform repeated experimentation with tuning, training, benchmarking, and deploying models continuously. This process requires frameworks like MLflow to manage the end-to-end process.

AI Explainability

Explainability is another emerging area that involves the reasoning of how the AI algorithms are making decisions. It involves inspecting how the AI decisions variables influence the outcomes and the steps involved in making the decisions. Model Explainability is becoming important as managers are looking for opportunities to eliminate any chances of bias, variability, or errors in the outcomes. For Instance, when using AI algorithm to predict the likelihood of losing a customer, business managers would like to understand the rationality behind the decision-making and check whether the model works consistently without any biases.

Diverse Skill Sets

A good AI application development involves leveraging the skills of multiple teams ranging from data scientists, application developers, infra architects, embedded developers (if edge inferencing is involved) along with domain experts. AI application development teams also need to have skills with benchmarking tools like ML workbench, DAWN, CPU/GPU benchmarking tools and so on, to look for constant improvement in performance. Finally, team shall be able to demonstrate explainability of algorithms to

business users with non-technical skills. All these frameworks continually evolve and getting the right skill mix is key to AI performance engineering and management.

Startups & Ecosystem Partners:

AI development due to its complex nature will require support from several stakeholders. Leading chip vendors are constantly improving their hardware, libraries, and development platforms. Startups are building new applications, custom models, frameworks, and accelerators for AI application development.

Given the complex combination of skills required, organizations require strong partnerships to connect missing links. For example, while building an AI application for diabetic retinopathy, it would be good to collaborate with the local ophthalmology institutes, doctors, and community volunteers, otherwise the model would not serve the purpose due to geography related drift issues and may not be useful for real-world scenarios in that region.

There is an increasing trend of niche startups building Industry-specific models with good accuracies and reliability. These models can be readily consumed in the overall application development.

AI Performance Management Team (AIPM)

Every AI application has its own complexity. It is good to have an AIPM team that brings the knowledge to deal with the complexities of AI application development, build reference architectures, frameworks, tools, and governance procedures.

AIPM teams can help navigate the complexities of AI application development by bringing diverse skills, and best practices rapid and successful deployment of AI applications.

The way forward

Organizations building enterprise-grade AI applications should rely on AIPM teams for advice on accelerated application development to cut down the risks and associated complexities.

About The Author

Sridhar CV

Sridhar CV heads Alliances for the CTO Incubation Team. He has over 20+ years of experience in the areas of Automotive Electronics, Railway Signaling systems, Telecom Networks and Governments Solutions, delivering multi-million-dollar technology transformation initiatives. Sridhar brings a blend of business and technology skills in the areas of AI & ML, drones, robotics and embedded systems, to accelerate innovation with the help of alliances, startups and other ecosystem partners. He represents TCS on the BIS LITD 27 sub-committee for IoT standards. Sridhar holds a Bachelor's in Electronics & Communication Engineering from JNTU Hyderabad and holds an Executive PGP in Strategic Management from IIM Indore.

Ravindran Subbiah

Ravindran Subbiah is an Entrepreneur-in-Residence (EIR) and heads the AI Performance Management Program at TCS. His vision is to industrialize AI/ML based solutions with processes and intelligent automation that provide enterprises a way to meet the increased expectations on AI-based solutions. Subbiah has used his 25+ years of experience across several organizations to develop effective value streams to operationalize AI-based solutions.

Nitin Hanjankar

Nitin Hanjankar is the Presales, Marketing and Alliances Head for the Research & Innovation (R&I) Incubation group at TCS. His presales responsibilities cover incubation programs such as Drones, Industry Operations, Cognitive and Rapid Labs. He has over three decades of industry experience. During his tenure, TCS has bagged multi-million-dollar, multi-year wins with leading multinational organizations, resulting in significant intellectual property (IP) revenue for TCS.

Experience certainty. IT Services
Business Solutions
Consulting

Contact

Visit the https://www.tcs.com/tcs-incubation page on https://www.tcs.com

Email: tcs.incubation@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that has been partnering with many of the world's largest businesses in their transformation journeys for over 50 years. TCS offers a consulting-led, cognitive powered, integrated portfolio of business, technology and engineering services and solutions. This is delivered through its unique Location Independent Agile™ delivery model, recognized as a benchmark of excellence in software development.

A part of the Tata group, India's largest multinational business group, TCS has over 453,000 of the world's best-trained consultants in 46 countries. The company generated consolidated revenues of US $22 billion in the fiscal year ended March 31, 2020, and is listed on the BSE (formerly Bombay Stock Exchange) and the NSE (National Stock Exchange) in India. TCS' proactive stance on climate change and award-winning work with communities across the world have earned it a place in leading sustainability indices such as the Dow Jones Sustainability Index (DJSI), MSCI Global Sustainability Index and the FTSE4Good Emerging Index.

For more information, visit us at **www.tcs.com**

All content / information present here is the exclusive property of Tata Consultancy Services Limited (TCS). The content / information contained here is correct at the time of publishing. No material from here may be copied, modified, reproduced, republished, uploaded, transmitted, posted or distributed in any form without prior written permission from TCS. Unauthorized use of the content / information appearing here may violate copyright, trademark and other applicable laws, and could result in criminal or civil penalties.
Copyright © 2021 Tata Consultancy Services Limited

Corporate Marketing | Design Services | M | 05 | 21