GIGAOM

REPORT

Delivering on the Vision of MLOps

A maturity-based approach

WILLIAM MCKNIGHT

SPONSORED BY Microsoft

GIGAOM

Delivering on the Vision of MLOps

A maturity-based approach

1. Summary

This report is targeted at Business and IT decision-makers as they look to implement MLOps, which is an approach to deliver Machine Learning- (ML-) based innovation projects. As well as describing how to address the impact of ML across the development cycle, it presents an approach based on maturity levels such that the organization can build on existing progress.

The paper is for practitioners who require practical advice on MLOps, rather than general principles of ML. It references ML-related activities and their importance but does not go into technical detail about the specifics of ML nor associated activities.

While the paper uses Azure Machine Learning and its documentation as a reference point, its guidance will apply to any ML environment.

Drivers to MLOps

Over the decades, data has proven to be a competitive differentiator. Once it was exclusively reports built by IT from overnight-batch-loaded data warehouses, but top performers have moved from passive reporting to predictive and prescriptive analytics, growing their skills in data science, and changing accepted paradigms as they derive insights to drive their businesses forward.

In recent years, rapidly falling costs of processing and increased throughput have unlocked new opportunities for organizations to maximize their data assets. Many companies have spent years or even decades collecting data in their data warehouses, data marts, data lakes, operational hubs, etc., and some now have the infrastructure and tools to act on the data optimally.

Based on established scientific principles, machine learning (ML) can deliver even greater levels of insight from data than traditional approaches, straight to the point of need, and without manual intervention. As shown in Figure 1, ML unlocks a broad range of opportunities across vertical sectors: the rewards will be greatest for those who have the skills, experience, and capabilities they need to deliver on its potential.

Figure 1. Use Cases are Becoming Vast for ML Pioneers Across a Wide Variety of Areas

	FLOW OPTIMIZATION	MODELLING AND ANALYTICS	PREDICTIVE INSIGHTS	THREAT AND RISK ANALYSIS
Public Sector	Traffic flow management	Smart city planning	Autonomous routing	Situational Awareness
Oil and Gas	Pipeline modelling	Drilling patterns and asset utilization	Intelligent planning	Safety assurance
Manufacturing	Supply chain optimization	Production optimization	Predictive maintenance	Fault identification
Retail	Supply chain optimization	Customer experience	Segmentation analysis and forecasting	Fraud and theft identification
Healthcare	Patient care pathway optimization	Disease research and drug creation	Early diagnosis of conditions	Patient safety
Technology	Operational efficiency	Log analysis	Capacity planning	Cybersecurity and zero-day detection

GIGAOM

These are still early days for ML, and success is not a given. Adoption faces several challenges:

- Senior management does not always see ML as strategic, and it can be difficult to measure and manage the value of ML projects.
- ML initiatives can work in isolation from each other, resulting in difficulties aligning workflows between ML and other teams.
- To be effective, ML training requires large quantities of high-quality data, which creates significant overheads across data access, preparation, and ongoing management.
- ML/data science work requires a large amount of trial and error, making it hard to plan the time required to complete a project.

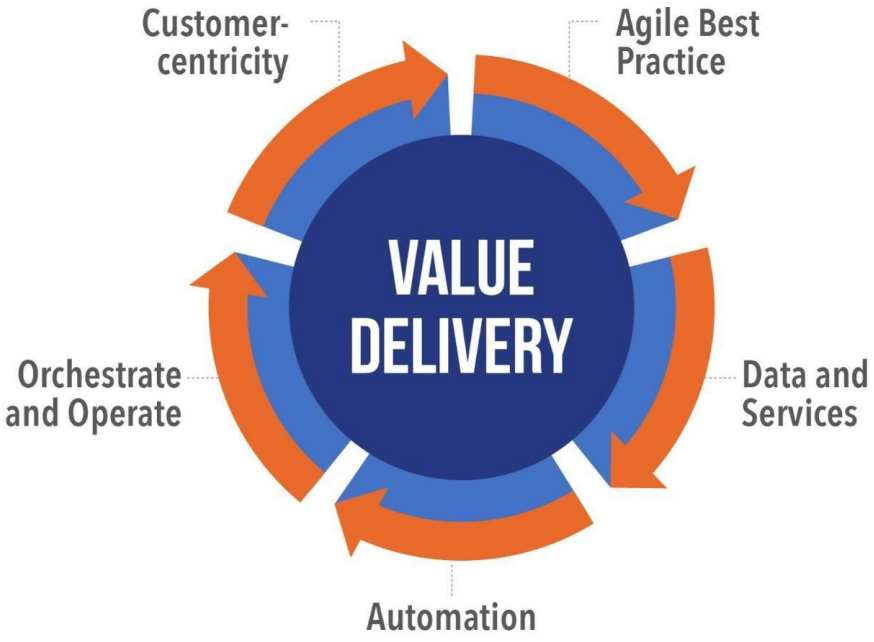
In response, ML adoption requires a cultural shift and a technology environment with people, processes, and platforms operating in the responsive, agile way organizations are looking to operate today: an approach we can call MLOps. Creating such a culture and environment cannot happen overnight: it comes by learning from those at the vanguard of ML how to map the potential of MLOps- driven innovation against an organization's specific needs and resources.

Based on multiple interviews with those working at the front lines of ML, a pattern emerges that sees the journey to MLOps success in terms of maturity. In this paper we present how to take MLOps strategy to practical reality, using best-practice principles and a maturity model that helps decision- makers assess, define, and enact steps towards MLOps leadership.

2. How Does MLOps Benefit ML?

MLOps draws on DevOps principles and practices. Built upon notions of work efficiency, continuous integration, delivery, and deployment, DevOps responds to the needs of the agile business – in short, to be able to deliver innovation at scale. To understand how to deliver MLOps, we need to consider both the function of DevOps, and how it is evolving.

Figure 2. The Goal of DevOps is to Assure the Delivery of Value to the Business, its Customers, & Other Stakeholders.



GIGAOM

As we cover in the GigaOm report Scaling DevOps: Strategy & Technical Considerations for Successful Enterprise DevOps, DevOps marks a cultural shift from slower, linear practices to agile approaches that bring in rapid iteration and parallelism, enabling developers to create and deploy innovative, software-based solutions. Since it was first instigated a decade ago, its

core ideas have not been standing still. Alongside such agile best practices, both enterprise requirements and infrastructure evolution (see Figure 2):

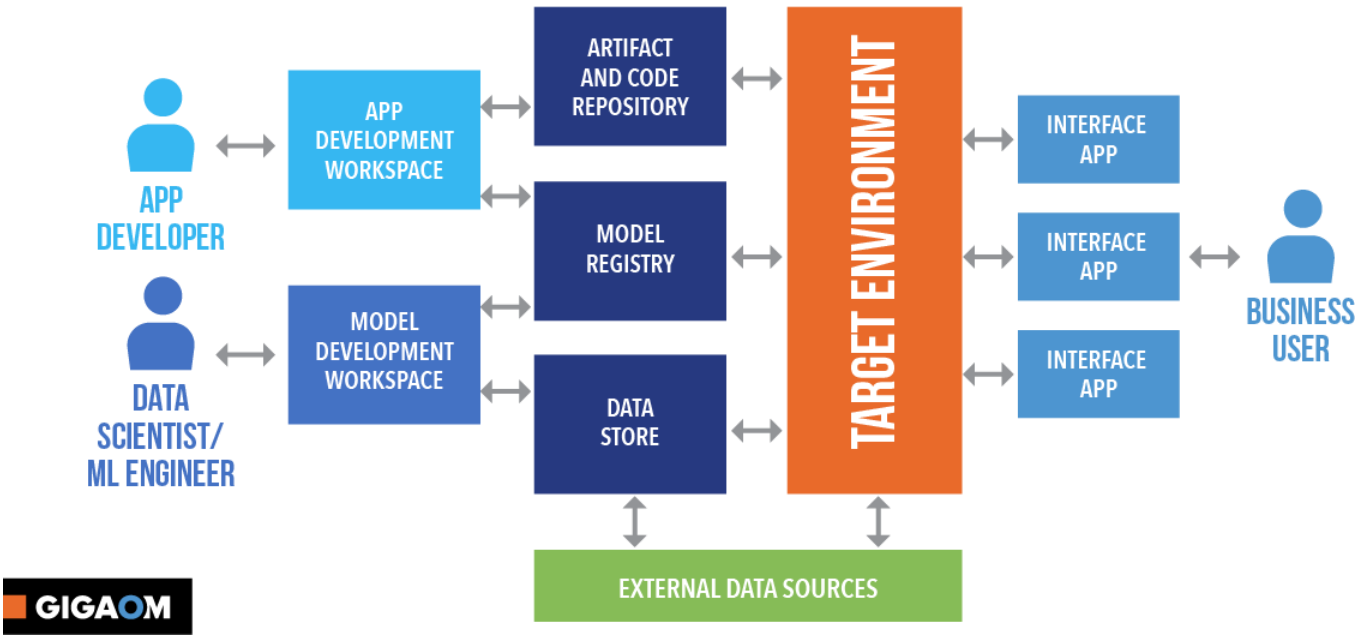
- **Shift to Customer-Centricity –** Today's business success stories are no longer about brand, product, selection, or model, but about how customers can achieve their goals.
- **Connect Data and Services –** DevOps success depends on how well platforms of data and existing/new services can be integrated, adapting to changing circumstances.
- **Automation –** Automation needs to be considered in the context of the above, to ensure constant, consistent, and efficient delivery of business value.
- **Manage Infrastructure Resources –** Applications will be deployed to an increasingly commoditized, flexible, target environment of infrastructure and platform-level services.

MLOps applies these principles to ML delivery, enabling the delivery of ML-based innovation at scale to result in:

- Faster time to market of ML-based solutions
- More rapid rate of experimentation, driving innovation
- Assurance of quality, trustworthiness and ethical AI

The MLOps process primarily revolves around data scientists, ML engineers, and app developers creating, training, and deploying models on prepared data sets (see Figure 3). Once trained and validated, models are deployed into an application environment that can deal with large quantities of (often streamed) data, to enable insights to be derived.

Figure 3. A Collaborative ML Environment Enables Data Scientists, ML Engineers, & App Developers to Share Resources & Innovate Together



Clearly, development of such models requires an iterative approach so the domain can be better understood, and the models improved. It also then needs automated tools, repositories to store and keep track of models, code, data lineage, and a target environment for deployment at speed to deliver an ML-enabled application. MLOps requires data scientists, ML engineers, and data engineers to work alongside software developers, so it can be seen as an extension of DevOps to encompass the data and models used for ML.

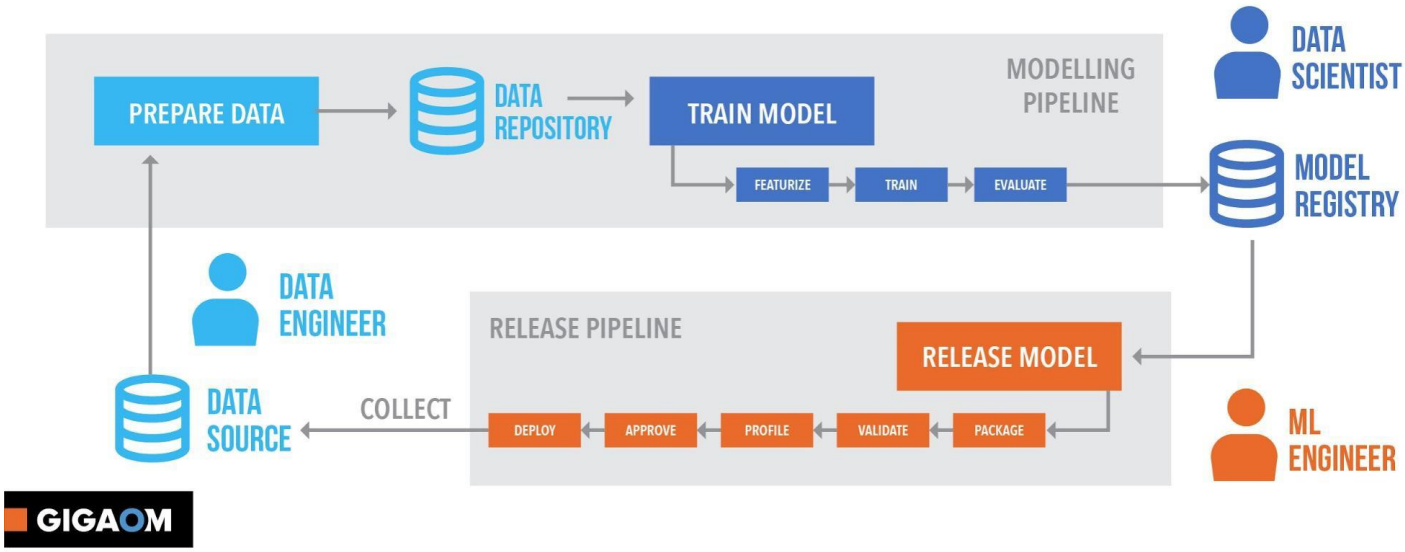
Terminology

MLOps uses terminology that draws from both DevOps and ML. In this document we refer to the following, as shown in Figure 4:

- **Pipeline –** An ML-based application will follow a planned and automated series of steps. The pipeline itself can be put under configuration control, such that the steps can be reused.
- **Datasets store/Datasets –** MLOps relies on an easily accessible and scalable source of data, both during training and inference. While data may come from several sources, it will be prepared, cleaned, and accessed as a single resource.

- **Repository –** A common, version-controlled storage resource (e.g. Git) for data, model, and configuration schemas, managing dependencies between models, libraries, and other resources.
- **Model Registry –** A logical picture of all elements required to support a given ML model, across its development and operational pipeline.
- **Workspace –** Model and application developers conduct their activities collaboratively using shared workspaces, accessible graphically or via code (e.g. written in Python), with access control over data sets, models, and insights.
- **Target Environment –** A deployment environment for ML models and code, packaged, for example, as containers/microservices, which is often cloud-based but can include on-premises and edge-based environments.
- **Experiment –** An activity sequence that enables a hypothesis to be tested and validated iteratively. Outputs of a given iteration need to be stored so they can be assessed, compared, and monitored for audit purposes.
- **Model –** Packaged output of an experiment that can be used to predict values or built on top of (via transfer learning).

Figure 4. The ML System Incorporates a Systematic & Repeatable Workflow.



3. Applying MLOps in Practice

To understand how this needs to look in practice, decision-makers can first consider the activities involved in the development of an ML-based application. These require data scientists working alongside application developers, following activities such as:

- **Collect/Prepare data** – Set up how data is ingested, prepared, and stored in a data repository.
- **Configure Target –** Set up the compute targets on which models will be trained.
- **Train Model –** Develop ML training scripts and submit them to the compute target for building and evaluating the models.
- **Register Model –** After a satisfactory run, store the persisted model in a model registry.
- **Release Model –** Validate results and if the model is satisfactory, deploy it into the target environment.
- **Operate Model –** Operate the deployed model in production, monitoring the model for inferencing performance and accuracy, data drifts, fault alerts.

To adopt DevOps terminology, this sequence of activities is called a pipeline. Activities in the pipeline are highly iterative – models need to be tuned, results tested, and data sources and models improved (See Figure 4). For example, engineers may discover that the insights that they need are associated with only a subset of the data sample; or some inherent bias exists in the results that needs to be addressed through additional data or improved algorithms; or discrepancies emerge between training and inference data sets – an issue known as data drift.

Successful ML teams respond to these challenges and deliver results by implementing the following best practices:

- **Reproducibility** – as with software configuration management and continuous integration, ML pipelines and steps, together with their data sources, code, models, libraries, and SDKs, need to be versioned and maintained such that they can be reproduced exactly as previously.
- **Reusability** – to fit with principles of continuous delivery, the pipeline needs to be able to package and deliver models and code consistently into training and target environments, such that the same configuration can be repeated with the same results.
- **Manageability** – the ability to apply governance, tracking changes to models and code throughout the development lifecycle, project tracking (for example through sprints), and enabling managers to measure and oversee both progress and value delivery.

- **Automation** – as with DevOps, continuous integration and delivery require automation to assure rapid and repeatable pipelines, particularly when these are augmented by governance and testing (which can otherwise create a bottleneck).

Through a shared approach, developers and data scientists can employ MLOps to collaborate and ensure ML initiatives are aligned with broader software delivery and more broadly still, IT-business alignment. Participants can adopt a test and learn mindset, improving outcomes while retaining control and assuring continued delivery of value over time.

4. A MLOps Scenario: Customer Churn

To illustrate these principles, let us consider a scenario. An online mobile phone retailer has been looking at how it can reduce churn across its customer base, that is, the length of time that customers stay loyal, particularly at the end of a subscription period. The principles are well-established: to reduce churn, the retailer needs to:

- increase customer loyalty and trust
- ensure that products and services are a good fit
- deliver targeted and effective marketing.

In practice, however, the linkage between these factors and customer behaviors is difficult to establish, and so requires an iterative, experiment-driven approach. The retailer has already had some success with ML and already has 5 data scientists and 5 experienced DevOps engineers. So, what needs to be in place to expand to an MLOps approach? First, this requires setting up the model and data environment; and second, pipelines are needed for training and inference. These activities are described below, taking into account reproducibility, reusability, manageability, and automation.

Configure Model and Data Environment

A first step is to prepare the ground for both data and modeling, by putting in place an environment that can manage both within an iterative process. Platforms and tools need to be able to support deployment to a wide range of target infrastructure and libraries optimized for ML, with multiple local/ cloud-based targets depending on model status.

Next, you need to identify and configure the data sources to be used by ML models. Activities include:

- Create and configure data stores, in this case, CRM data
- Normalize, transform and otherwise prepare datasets for training and inference

- Point algorithms and code to the data
- Enforce transparency (e.g. through audit trails) to build confidence in results

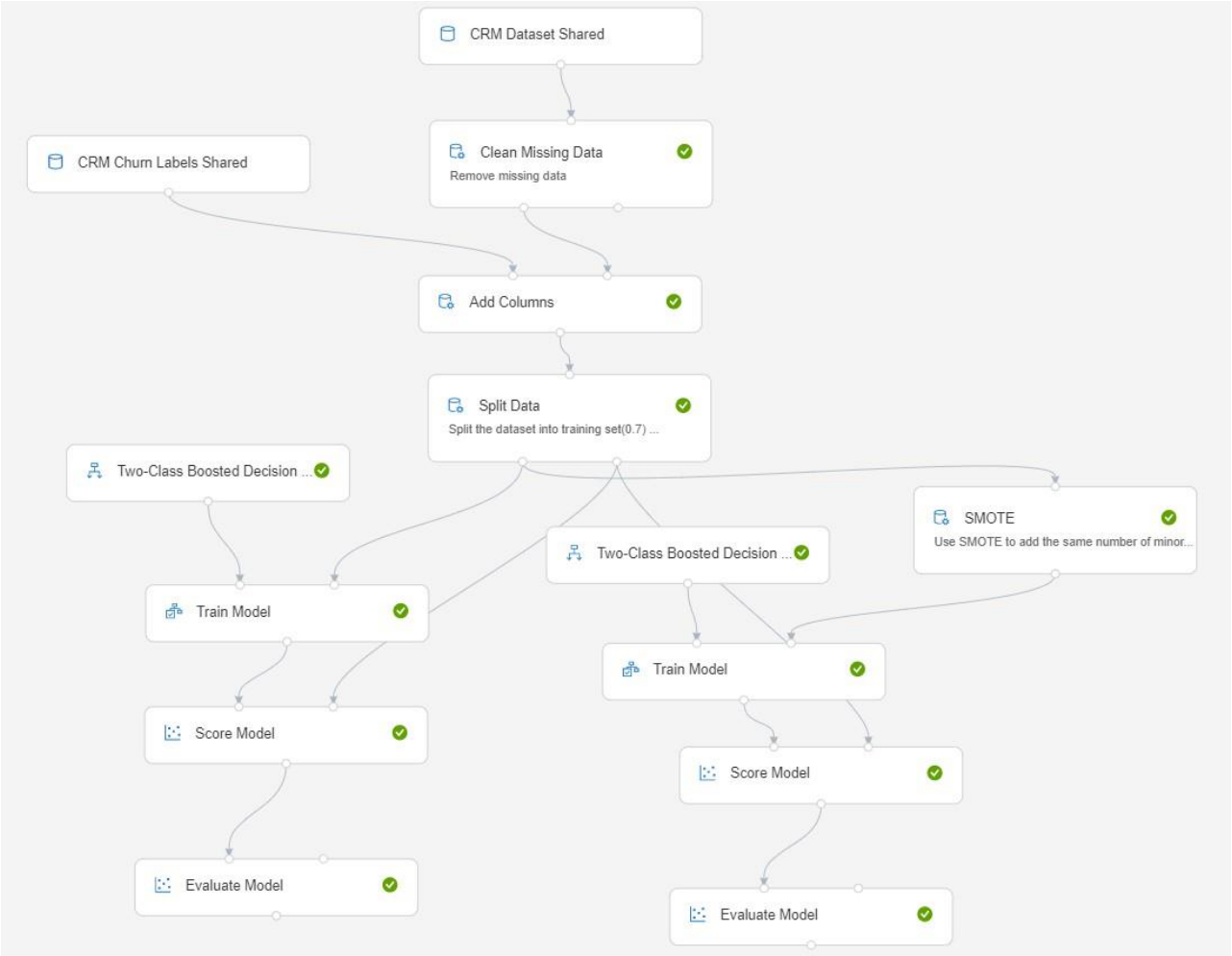
Note that pipeline steps might also consume data sources and produce “intermediate” data, which is then available for other steps later in the pipeline. Pipeline steps can also be reused. So, for long- running data preparation jobs you can use the output of this particular step to feed several other steps or training in parallel and save on execution time.

In this scenario, customer loyalty factors may be fed back into the model as variables, for example, testing the effectiveness of historical loyalty schemes.

Create Pipelines for Training and Inference

With an environment and data in place, it is possible to consider how to organize the flow of model creation activities from training, through validation and testing, to operation and inference, as reproducible pipelines. During training and other steps, scripts can read from or write to datastore, and records of execution are saved as runs in the workspace, grouped under experiments. An example pipeline is shown in Figure 5.

Figure 5. Pipeline Steps Link Data Preparation, Model Training, & Evaluation.



By reviewing these experiments, the retailer can monitor results for applicability and effectiveness of insights. Speed and repeatability are key to this test-and-learn approach, requiring both pipeline reproducibility and automation. At the same time, data scientists need to ensure that the models deliver the right prediction; for example, assessing whether sample data has been subject to data drift between training and inference, which would impact the quality of results (see Figure 6).

Figure 6. Evaluation Tools Enable Issues Such as Data Drift to be Assessed.

A complete demonstration for this churn example is available here. This shows how you can use the Azure Machine Learning SDK to create and publish a pipeline into an Azure Machine Learning workspace.

5. The MLOps Maturity Model

> *Successful ML teams can address the challenges of developing AI solutions by applying MLOps and implementing the practices of reproducibility, automation and manageability. Organizations should also routinely measure progress and understand where they are at each stage for sustained success and the MLOps maturity model from GigaOm can be a valuable tool in this journey.*
>
> *–Vijaya Sekhar Chennupati, Applied Data Scientist/Data Engineer, Johnson Controls*

As data scientists iterate more quickly through test-and-learn cycles enabled by MLOps, they can arrive at genuine insights more quickly, however, this cannot be at the expense of compromising quality or governance. Therefore, to deliver ML in a way that delivers on the goals of MLOps, enterprises need a way to measure their progress. Maturity models help

organizations understand where they are on the journey and what steps they can undertake to "level up."

Levels in the MLOps maturity model have been defined based on enterprises on or yet to start the ML journey. As a result, five nominal levels are identified. While the distribution of companies interviewed was highly skewed to the lower end, this should be of no consequence to the leading organizations aspiring to sustained success. Indeed, it represents an opportunity for faster movers.

In the Maturity Model, categories of response exist across:

- **Strategy** – how well a company can align MLOps activities with executive priorities, organizationally and culturally
- **Architecture** – the ability to manage data, models, deployment environments, and other artifacts as a coherent whole
- **Modeling** – data science skills and experience, enabling domain knowledge and the delivery of models that correctly represent a domain
- **Processes** – efficient, effective, and measurable delivery and deployment of activities, across scientists, engineers, and operational management
- **Governance** – overall, the ability to build secure, responsible, and fair AI solutions, to place trust in inputs, outputs, and explainability of ML models

In general terms, maturity levels tend to move in harmony across these categories and attaining and retaining momentum up the model is paramount for success. As a general principle of maturity models, an organization cannot skip levels in any category nor advance in one category well beyond the others.

Vis-à-vis MLOps, organizations will ascend the model through concerted efforts delivering business wins utilizing progressive elements of the model, thereby increasing their machine learning maturity. The model should give a sense of the priority of each element and how to roadmap efforts. No two organizations will follow the same route; however, the model provides goals, markers, and activities that are relevant for all organizations.

6. Delivering on MLOps Maturity

An aspirational goal is to reach a stage in which MLOps is proactively adopted across ML teams and initiatives. So, how can an organization deliver? Each maturity level has certain characteristics and relevant actions to build to the next level. For example, most organizations start with a singular project with minimal staff, expanding across other applications and business groups.

Level 0

At this level, organizations are still to accept a place for ML, let alone MLOps. They have no plans to employ data scientists and little understanding of the complexity they will face. Processes are manual, with little or no ability to measure success.

CATEGORY	CURRENT STATUS
STRATEGY	– No data scientists hired – Sceptical of value of ML among executive team
ARCHITECTURE	– Data Silos with one-off integration – Data not prepared nor ready for ML
MODELING	– Manual process for model training – Limited pilot studies
PROCESSES	– No DevOps practices adopted – No clearly defined success criteria for ML projects
GOVERNANCE	– Not considered

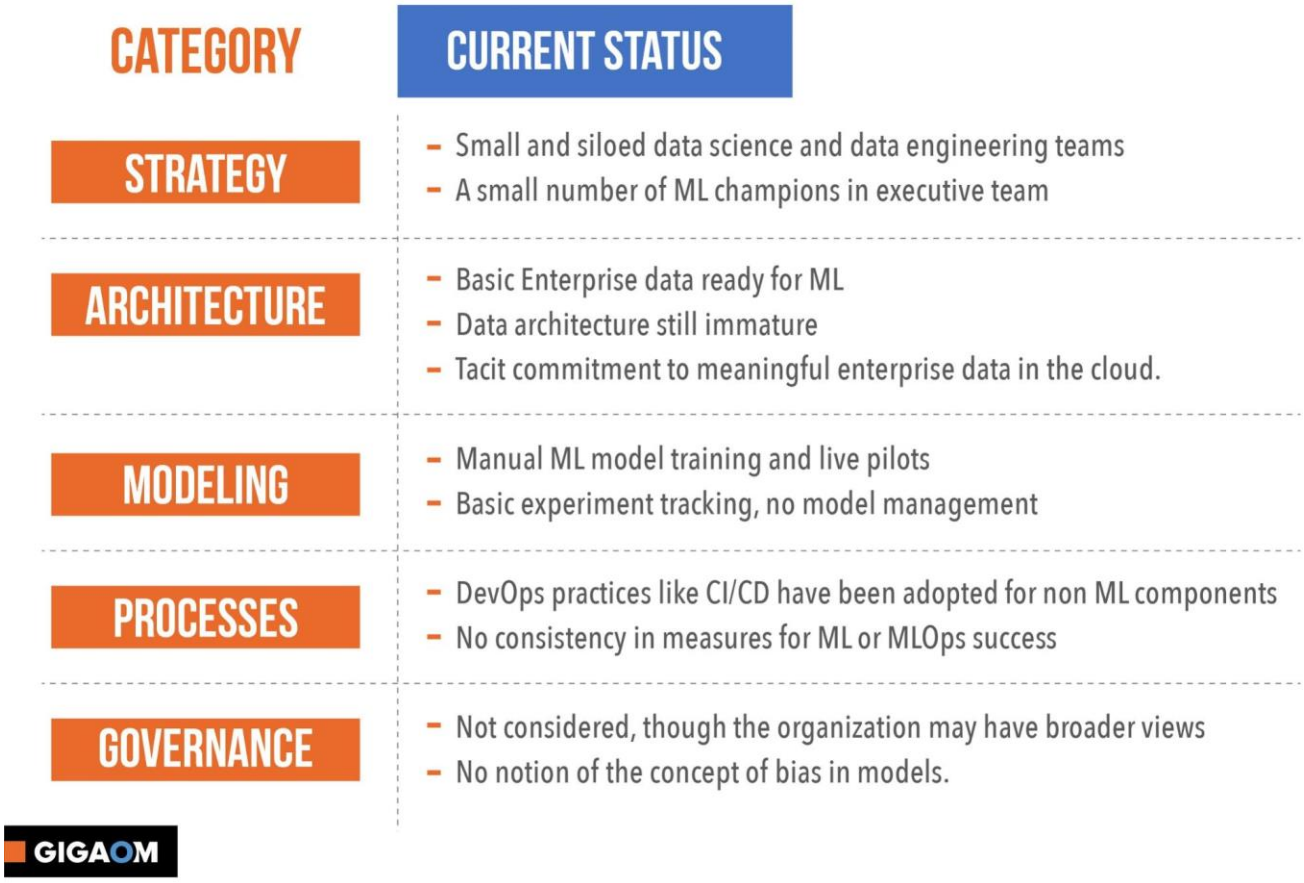
The priority at this stage is therefore to accept the need to build data science into business or technology strategy. It is also to start to get data into shape for better analytics, and therefore ML, which means data is:

- Of a suitable level of granularity and at a data quality standard
- In an appropriate and scalable platform for its profile and usage
- Accessible to multiple business groups

The target outcome at this level is to offer a basis for ML, and therefore MLOps, as a driver of tangible business value. One activity could be a brief pilot that can demonstrate the value held within existing data, and therefore the opportunity created by ML, at the same time as kicking off skills development.

Level 1

These organizations tend to be good at traditional data capture and analytics and may have a level of commitment to cloud-based approaches; for example, they may already be accessing data from the cloud. However, they are yet to engage with ML in a strategic way.



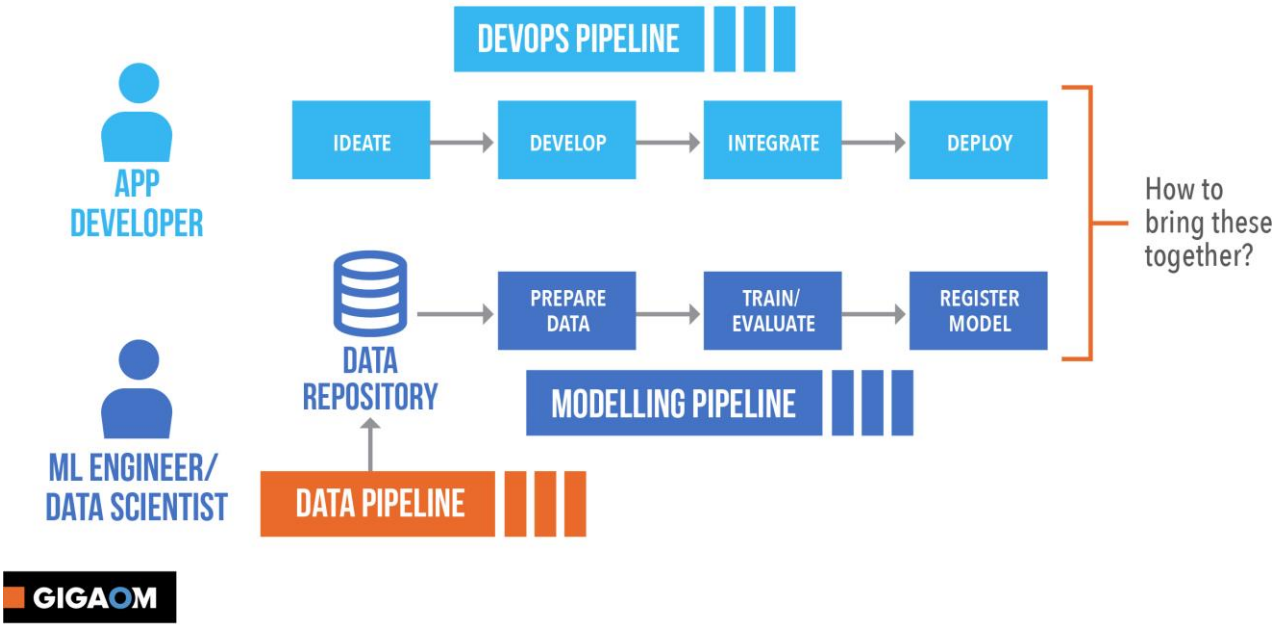
There will be only pockets of data science skills on the team and limited experience of what benefits ML can bring. The main challenge for this level is to know how to kick off ML activities meaningfully, based on existing understanding of DevOps and what it can bring to the table.

The priority at this stage is to start building skills and expertise, together with sufficient buy-in at the right level. We would recommend the following actions:

- Focus on developing an architectural understanding, in terms of the kinds of common architectures that may be applicable.
- Get in shape for operational aspects by bringing in collaborative and agile practice across both development and ML activities.
- Identify and set out key success measures, as part of ML strategy.

- Get data assets under management.

Figure 7. At Level 1, Development & Model Creation Happen Independently.



In terms of outcomes, recall the key MLOps criteria of reproducibility, reusability, manageability, and automation. Level 1 focus can be on improving reusability of software and model artifacts, and delivering automation by putting the right environment in place, creating a pipeline, then using it to prepare data and train a model. The overall goal, alongside building initial skills and exploring possibilities, is to generate a model that demonstrates the value of ML.

As this GigaOm report covers, some organizations start with more automated use of ML to prove the use case, before moving onto integrating with applications with MLOps.

Level 2

Manager organizations are actively looking to deliver the benefits of ML and have made some discrete progress but now need to consolidate and co-ordinate so their efforts can scale. MLOps allows your efforts to scale. “There is no better place to apply Azure Machine Learning’s MLOps than Office products due to the very large scale (hundreds of millions of users daily) of the product. MLOps was intended for this level of big data,” said Anand from Microsoft Office, whose story can be found below.

CATEGORY	CURRENT STATUS
STRATEGY	– Small Data science, data engineers and software development teams starting to be coordinated – ML development efforts still unstructured and discrete.
ARCHITECTURE	– Data architecture is mature – Most enterprise data ready for ML in the cloud – Overt commitment to cloud
MODELING	– Experiment tracking and model management in place – Models dependencies not well understood
PROCESSES	– Development iterative but CI/CD not in place for models – ML infrastructure expertise not broadly available – ML configuration is an afterthought – Metrics/ measures in place but not consistent across projects.
GOVERNANCE	– Model explainability not considered – Models may harbor prediction bias – Model releases are tracked

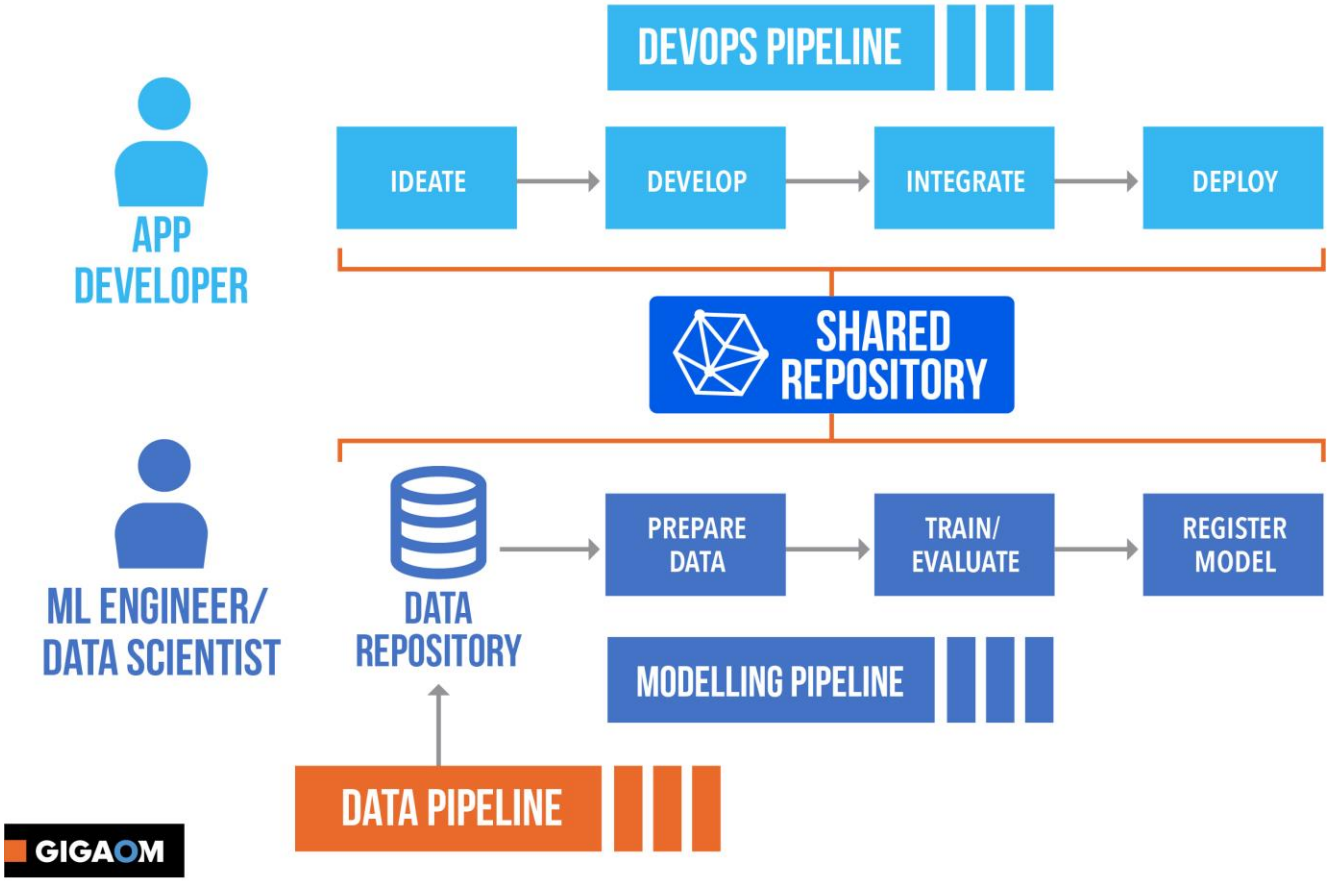
GIGAOM

Organizations at this level may be coordinating development and ML activities but are inefficient and struggle to deliver value in a measurable way. The opportunity exists to start

to drive MLOps as a practice, assuring a framework of governance and tooling that can minimize bottlenecks as efforts progress.

The pipeline becomes the center of gravity for models going to production. Data scientists may still be using tools such as Jupyter for their day to day operation, but the pipeline becomes the key in moving from exploratory to production. MLOps is what will bring it to production.

Figure 8. Level 2 Adds Reproducibility & Iteration to Development & Model Creation.



We would recommend a focus on the following:

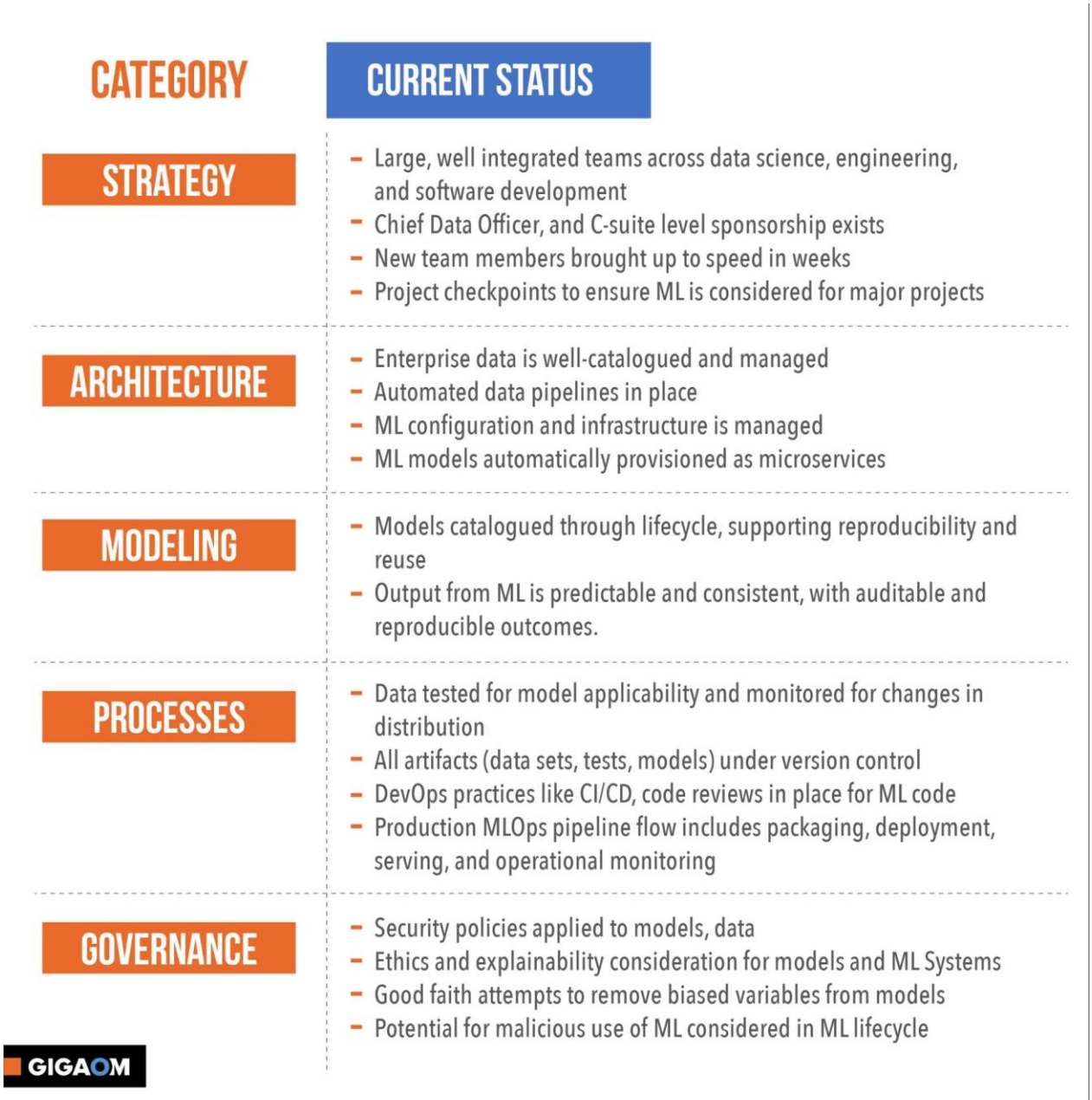
- Increase recruitment of data scientists, looking at training so new team members can be brought up to speed quickly.
- Acknowledge operational aspects – monitoring, validation, profiling, testing, drift and many of the issues that come up with automating, operationalizing and monitoring applications.
- Bring governance and ethical understanding into ML strategy, recognizing complications such as algorithmic bias and the importance of comprehensive data sets.

- Implement planning, management, and prioritization, standardizing measures and adopting a shared repository to track all models along their lifecycle.
- Review how ML can be tacked onto existing priorities, through an ML project identification process – particularly BI/analytics.

Expected outcomes for Level 2 organizations should build on reusability and automation, adding reproducibility such that multiple collaborators can work on the same models and data, with consistent results. The goal is to put as much as possible under version control, at the same time as improving collaboration as the number of data scientists increases. Equally, Level 2 can bring in manageability, such that models can be created consistently and at increasing velocity based on agreed measures.

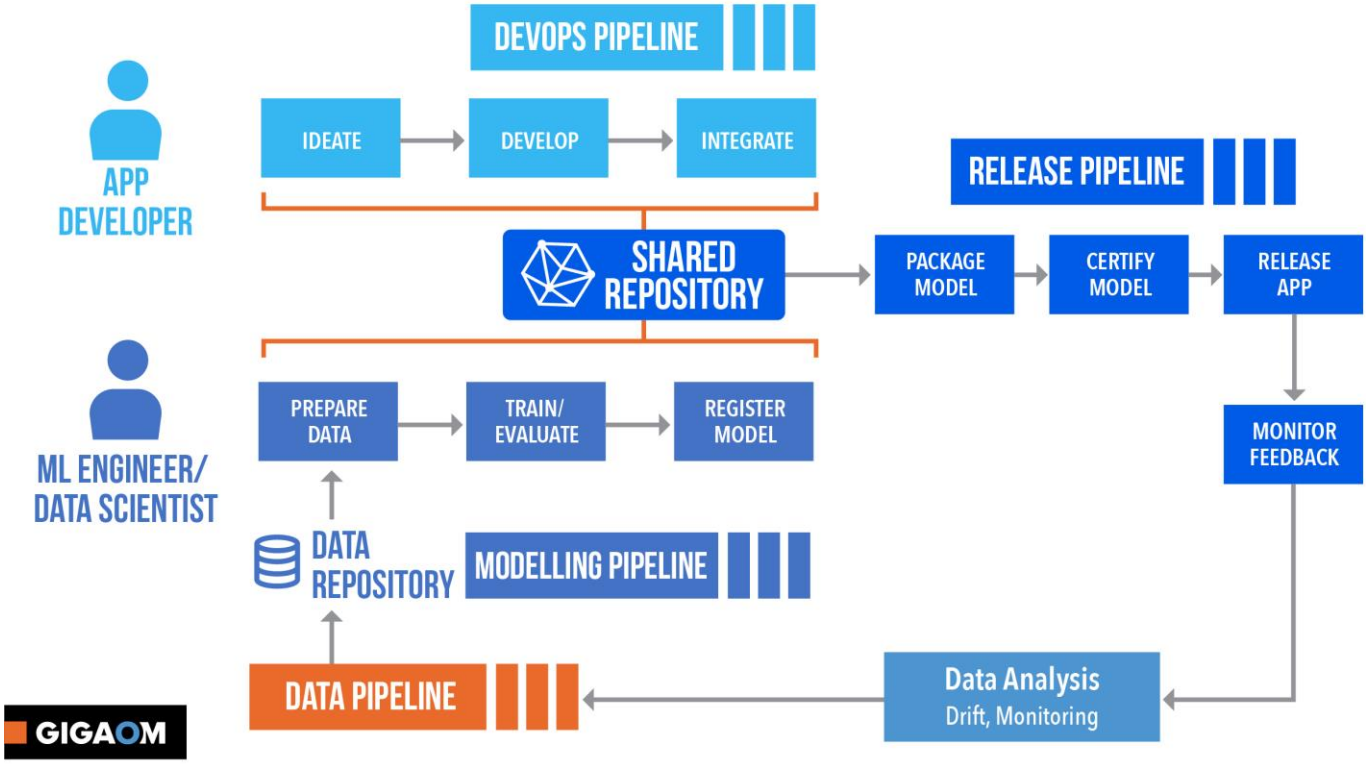
Level 3

At this level, organizations are delivering ML in an efficient, effective manner, have seen the benefits produced, and have grown to multiple data scientists. They have a program synergistic with MLOps and can start to look at how they optimize their efforts. AI developers are working closely with software developers in a workflow that is consistent across the MLOps and DevOps flows and processes.



Proactive adoption of MLOps means you have a tightly controlled lifecycle for your ML and a process for retraining models. Chandra Setiawan, Principal Software Engineer at Microsoft Office says, “That’s what the power of Azure Machine Learning MLOps gives us.”

Figure 9. At Level 3, DevOps & ML Pipelines Converge to Create a Shared Release Pipeline.



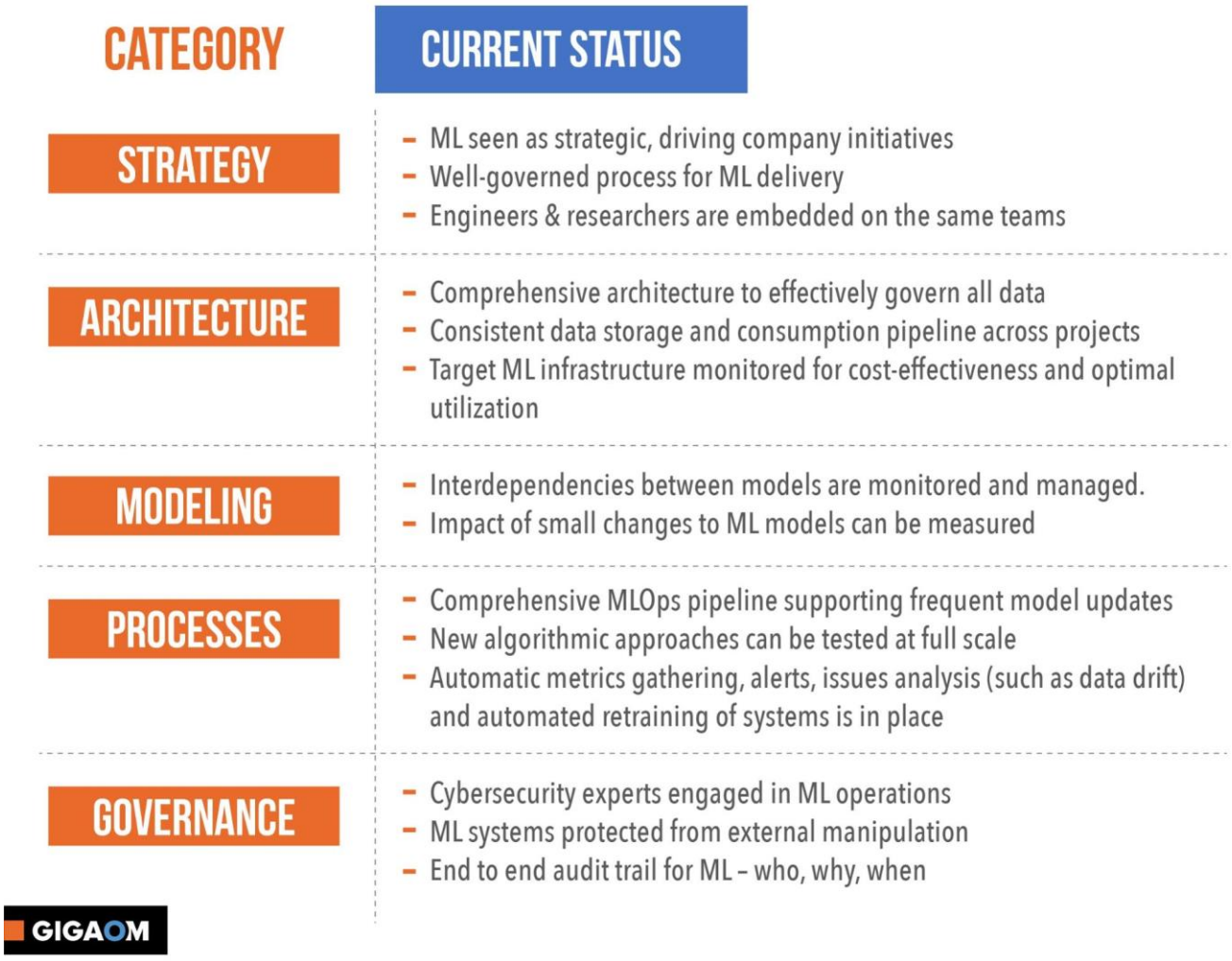
While the organization is delivering successfully, it can do more to broaden the use of ML across the organization. We recommend a focus on:

- Driving measures and value delivery, so business value can be clearly assessed and reported
- Closing the operational loop to feed MLOps into the model development cycle

At Level 3, organizations can focus on delivering across all four criteria, delivering reproducible outcomes on top of reusability, manageability, and automation. Success across all of these key criteria enables leveling up to a true MLOps leadership position, which offers a solid platform for scaling up ML efforts.

Level 4

For these organizations, the business is fundamentally different than it was prior to its MLOps maturity journey, due to the integrated use of ML. Governance has become a central plank of ML strategy, considering both the potential for malicious use of ML in the MLOps lifecycle and the need to ensure outcomes that are explainable and transparent.



Even successful ML leaders must ensure that they can sustain the use of ML, keeping up with technology and algorithmic advances, and responding to new opportunities. To assure continued improvement in MLOps, we recommend:

- Have an MLOps governance board to review progress and set direction.
- Share experiences with industry partners to further drive innovation

Level 4 should see an organization having scaled up ML efforts following a consistent approach, therefore in a position to broaden experimentation and use different techniques. Organizations will be able to assess performance and behavior, looking at questions such as data drift to further enhance and optimize ML-based applications.

Conclusion: Proactively Adopt MLOps

As seen, MLOps is an essential discipline for a winning organization – one that intends to retain relevancy through the next decade, through faster time to market, and trustworthy and ethical AI using platforms such as Azure for Machine Learning. The following summary recommendations can support the creation of an action plan, wherever an organization is in maturity terms:

Be prepared for ML-driven change. As Sze-Wan Ng Director, Analytics & Development at TransLink said, “The organization must have a willingness to try new things.” TransLink took risks with ML and it paid off. No plateau is comfortable for long, and organizational competitiveness of the future will be defined by ML and MLOps. ML-driven success will not come to organizations that keep the status quo.

Align the ML roadmap with business priorities. Budgets seldom arise specifically for improving maturity. The skilled enterprise technology leader manages potentially conflicting goals, delivering both business wins and improved maturity. The MLOps leader is in a prime position to not only deliver business wins but also create the projects that utilize machine learning. Infusing ML into the enterprise’s endeavors is as important as creating new projects for ML.

Focus on onboarding. Particularly lower-level organizations can recognize data as a discipline, not an afterthought. When new data science team members are brought into the organization, they can be brought up to speed in weeks, not quarters, based on documentation, adherence to reasonable, common constructs in the data environment, and defined business goals.

Embrace transparency and predictability. For MLOps success, incorporate governance and security frameworks from the outset. For example, though they may not have perfected it yet, ML Operators take measures to eliminate model bias and the potential for malicious use of ML. They understand the importance of re-running experiments to replicate results and have overcome difficulties in selecting the correct ML algorithms.

Overall, organizations can target fast-track movement toward active adoption of MLOps, incorporating all of the elements stated above and with an eye on the next level, within the next one to three years. Enterprises that remain at a lower level will find themselves at a significant disadvantage compared to those in a position to scale their ML efforts to deliver real business advantage.

8. Real World Use Cases

TransLink

Company Background

TransLink is Metro Vancouver's transportation network, serving residents and visitors with an extensive bus system throughout the region, SkyTrain rapid transit, SeaBus passenger ferries, West Coast Express commuter rail, and HandyDART for passengers who are unable to use conventional transit.

Opportunity

TransLink is in the business of delivering reliable transit services and their customers depend heavily on accurate bus departure times to plan their journeys. With the expansion of bus services and growth in the region, their bus predictions were getting worse and less reliable. Customer complaints were mounting. TransLink needed to come up with a better solution.

Challenges

Transit customers want more precise departure times for their stop. Yet so many things can impact a predicted time. Traffic. Bad weather. Disruptions.

Based on their experience with the Microsoft Garage project, where they had used ML to successfully predict bus crowding, they decided to reach out to Microsoft to explore ML and MLOps to generate stronger prediction models.

Solution

The project started with a proof of concept to experiment on the use of ML. A simple model was built based on six weeks of training data and the results were promising enough to continue to a pilot of 13 routes. These routes were carefully selected to stress test the robustness of the model under a wide range of conditions.

A modeling approach was used to drive localized information into each component to produce the most accurate stop-level predictions using real-time bus location and road condition data. The models were trained using two years of historical data. The ML model chosen was Extreme Gradient Boosting (XGBoost) because of its fast training time and high-performance rate.

Each bus stop and segment in the transit system has its own set of machine learning models, provided that there is enough data to train on (at least 500 instances of dwell/run events). Consequently, they have over 18,000 different sets of models. MLOps was used to deliver a

training pipeline that could manage these models, enabling the creation of a model across all the bus stations involved.

As the project progressed, the biggest influence on model accuracy was found to be location of the bus and how long the buses sit at their stops: the results were found to be better when not influenced by the schedule.

All success criteria exceeded expectations. The difference between predicted and actual bus time dropped by 74%. The average customer spent 50% less time waiting at the stop and the number of customers waiting more than five minutes (“in the dark”) dropped by 10%.

The company is enthused by the results and looking to expand ML. Participation and interest was high. However, Sze-Wan Ng, the Director of Analytics & Development, shared that the Ops was harder than the ML, and they look forward to maturing their MLOps program with Azure MLOps, which will be necessary for the expansion of machine learning at TransLink.

Microsoft Office

Opportunity

Office has been on a two-year journey to infuse AI in the Office products with the goal of amplifying human productivity. Anand Balachandran, Principal PM Manager in Office Language and Intelligence said that “time and human attention are the most valuable commodities in the world today,” and “with the power of AI we’re helping users save time and effort, providing them the right insights and thereby getting them to their outcomes quicker and making them more productive. We want to enable experiences where a feature would automatically be there when users need it; so users can spend their valuable time working on their business problems and getting their ideas together, rather than trying to hunt down the right feature or command in Office.”

Challenges

A team of applied data scientists work on training ML models for each of the AI features. One of the biggest pain points was taking the trained models and deploying them to a production environment. There were a lot of “pieces here and there,” according to Anand, but no clear end-to-end solution that did this seamlessly respecting all the compliance requirements for Office Cloud Services.

Solution

Anand noted that the selection of MLOps with Azure Machine Learning was primarily based on its effective ML lifecycle management – a scalable, faster, and automated way to get trained models into production.

Azure Machine Learning, with its inherent support for Notebooks, helps keep code and data together, and version and snapshot the data to automate the training process. The Office team started with a proof of concept of MLOps with Azure Machine Learning. With the ML training and toolkits they learned how "MLOps with Azure Machine Learning ties everything together from start to finish," according to Anand.
As models are deployed into production and more users experience the model recommendations, they generate ML feedback loops, which can trigger retraining of the models based on user behavior and make the models better.

As an example, the Editor team is using Azure Machine Learning to make Editor even smarter and ensure immediate response to errors, even when a user is working offline. Using advanced neural machine learning techniques, the scientists in the Editor team are training recurrent, convolutional, and transformer-based neural models that can better understand the linguistic intricacies in any language and detect errors even in text that is not grammatically conformant (such as headers). Since the models they train have millions of parameters, they need to employ distributed model training across a cluster with hundreds of GPUs. The team's workflow involves a small team of data scientists collaborating on a Git repository hosted by Azure DevOps.

Azure Machine Learning service integrates with Azure DevOps to simplify the machine learning lifecycle. The Editor team investigates the kinds of models that would provide real value to Office 365 users. After that, they use state-of-the-art technology such as the open-source PyTorch and Horovod, supported by the Azure Machine Learning service offering, to implement these features. Then the team tests the models both offline and online extensively until they are certain their features add value to the Office user experience.

ML algorithms used in Office don't need to be complicated to be successful. Simple linear regressions are a popular model utilized, though they also use deep learning where it is best, like when the model cares about long-tail data. For instance, in the case of command-prediction, they moved from a shallow model to a deep model, which can incorporate an extensive personal history in offering the suggestions. Deep learning is employed when they have the right amount of labeled data and the resultant models meets the performance requirements (memory, latency, etc.) that the scenario warrants.

For the Office team, MLOps with Azure Machine Learning is a one-stop solution for training, inferencing, and automated deployment, with low friction. Azure MLOps gives the Office team an auditable and repeatable process for model training and deployment.

Over the last two or more years, the team has been able to find many opportunities across the Office product for ML. They actually have an AI feature roadmap alongside the product roadmap. Anand says they are in a "generational shift into digital transformation of business, and we're constantly trying to amplify human productivity."

Every time you deploy a new build of Office, you will likely get new value, thanks to machine learning and the efficiencies made possible by MLOps with Azure Machine Learning.

Company Background

Microsoft (Nasdaq “MSFT” @microsoft) enables digital transformation for the era of an intelligent cloud and an intelligent edge. Its mission is to empower every person and every organization on the planet to achieve more.

Additional materials and examples of Microsoft’s Azure Machine Learning, and MLOps, are available as follows:

- Free trial: https://azure.microsoft.com/en-us/free/services/machine-learning/
- Website: https://azure.microsoft.com/en-us/services/machine-learning/mlops/
- GitHub: https://github.com/Microsoft/MLOps
- Documentation: https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-model- management-and-deployment

Johnson Controls

Company Background

For over 130 years, Johnson Controls has produced fire, HVAC, and security equipment for buildings. The company, with its broad array of product lines, is at the forefront of the smart city revolution, with data science and machine learning (ML) as key elements of their approach.

As part of its focus on digital transformation, Johnson Controls formed a data science team four years ago and now employs over 50 people involved in ML projects. The vision is to scale use of ML from small equipment and sensors to entire buildings. The company is currently involved in what will be the smartest building in the world, in Dubai.

Opportunity

To deliver on its vision, Johnson Controls needed to keep pace with the demand for its building equipment without scaling the cost. Chillers, in particular, are key to building efficiency. Johnson Controls runs over 4,000 chillers, streaming time-series data volumes up to several terabytes. The opportunity to improve both efficiency and function through real-time data analysis was abundantly clear.

Engineering teams had begun to gain insights from the chiller sensor reads but recognized the need to build fully predictive models. The company knew it needed to turn to the cloud and machine learning for new, improved insights that would deliver both cost savings and better business decision making.

Challenges

There were many challenges Johnson Controls faced in their ML journey before implementing MLOps. To start, the company was building its own ML pipelines without benefiting from external best practices. Most activities were manual, including version control and model deployment. As the scale of activities increased, so did the size and complexity of models needing to be stored.

Data scientists also recognized the need to store additional information, such as model weightings. Over time, data drift required models to be re-trained, creating a different set of weightings for the same architecture: this caused the need for a model registry, with standardized storage – and the bandwidth to manage it.

At the same time Johnson Controls knew they needed a proper, collaborative pipeline for model development, rather than the "black-box" approach they had been employing.

Solution

As they started to look beyond the organization and looked at MLOps as a principle, Johnson Controls reached out to Microsoft to ask what existed in terms of best practice. MLOps was a natural fit with Johnson Controls, given their existing use of DevOps, and they were able to transfer many of its good practices into their ML operation.

With the MLOps capabilities in Azure Machine Learning, they are now able to version and install models quickly. According to Vijaya Sekhar Chennupati, applied data scientist/data engineer at Johnson Controls, the value from MLOps is in the model registry. "We are able to containerize models automatically and deploy them. The ability to maintain and monitor model history is a huge productivity gain," he says. As a tip, Johnson Controls recommends MLOps for all new models, and when the value is understood and processes are tighter, to backtrack MLOps into all existing models.

In addition, the company was able to build models more collaboratively, using experiments to test and validate hypotheses against a model architecture, to test for improved results. Improved controls meant that new projects could be jump-started based on existing models and data sets. The cost of entry was reduced at the same time as the cost of management. "Ease of model versioning and dependency management makes our process more agile, helping traceability of deployments," says Chennupati.

Many of the Johnson Controls models are in MLOps today, which enables Chennupati and his team to conduct experiments that they otherwise would not be able to do, resulting in real business value. Integrating with existing DevOps is also easier. Explains Chennupati, "By partnering with Microsoft, we helped develop the MLOps that gets integrated into Azure DevOps, to create a seamless and integrated CI/CD process using Azure DevOps."

With regard to the chillers, utilizing the time series data off the sensors in real-time, Johnson Controls was able to develop models to optimize predictive maintenance routines. Because

data science teams now have the ability to process the data, Johnson Controls now has up to 70 different types of sensors on their chillers with more coming online.

Overall results are that both unplanned downtime and mean time to repair are down by two-thirds. Chiller shutdowns, a highly detrimental event for JCI, has been drastically reduced. They predict potential shutdowns several days before and can comfortably intervene to keep customers happy and save money and time. “Using the MLOps capabilities in Azure Machine Learning, we were able to increase productivity and enhance operations, going to production in a timely fashion and creating a repeatable process,” says Chennupati.

9. Annex: An MLOps Maturity Model

CATEGORY	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
STRATEGY	– No data scientists hired – Skeptical of value of ML among executive team	– Small and siloed data science and data engineering teams – A small number of ML champions in executive team	– Small Data science, data engineers and software development teams starting to be coordinated – ML development efforts still unstructured and discrete	– Large, well-integrated teams across data science, engineering and software development – Chief Data Officer, and C-suite level sponsorship exists – New team members brought up to speed in weeks – Project checkpoints to ensure ML is considered for major projects	– ML seen as strategic, driving company initiatives – Well-governed process for ML delivery – Engineers & researchers are embedded on the same team
ARCHITECTURE	– Data Silos with one-off integration – Data not prepared nor ready for ML	– Basic Enterprise data ready for ML – Data architecture still immature – Tacit commitment to meaningful enterprise data in the cloud.	– Data architecture is mature – Most enterprise data ready for ML in the cloud – Overt commitment to cloud	– Enterprise data is well cataloged and managed – Automated data pipelines in place – ML configuration and infrastructure is managed – ML models automatically provisioned as microservices	– Comprehensive architecture to effectively govern all data – Consistent data storage and consumption pipeline across projects – Target ML infrastructure monitored for cost-effectiveness and optimal utilization
MODELING	– Manual process for model training – Limited pilot studies	– Manual ML model training and live pilots – Basic experiment tracking, no model management	– Experiment tracking and model management in place – Models dependencies not well understood	– Models cataloged through lifecycle, supporting reproducibility and reuse – Output from ML is predictable and consistent, with auditable and reproducible outcomes	– Interdependencies between models are monitored and managed – Impact of small changes to ML models can be measured
PROCESSES	– No DevOps practices adopted – No clearly defined success criteria for ML projects	– DevOps practices like CI/CD have been adopted for non-ML components – No consistency in measures for ML or MLOps success	– Development iterative but CI/CD not in place for models – ML infrastructure expertise not broadly available – ML configuration is an afterthought – Metrics/ measures in place but not consistent across projects	– Data tested for model applicability and monitored for changes in distribution – All artifacts (data sets, tests, models) under version control – DevOps practices like CI/CD, code reviews in place for ML code – Production MLOps pipeline flow includes packaging, deployment, serving and operational monitoring	– Comprehensive MLOps pipeline supporting frequent model updates – New algorithmic approaches can be tested at full scale – Automatic metrics gathering, alerts, issues analysis (such as data drift) and automated retraining of systems is in place
GOVERNANCE	– Not considered	– Not considered, though the organization may have broader views – No notion of the concept of bias in models	– Model explainability not considered – Models may harbor prediction bias – Model releases are tracked	– Security policies applied to models, data – Ethics and explainability consideration for models and ML Systems – Good faith attempts to remove biased variables from models – Potential for malicious use of ML considered in ML lifecycle	– Cybersecurity experts engaged in ML operations – ML systems protected from external manipulation. – End to end audit trail for ML – who, why, when

10. About William McKnight

An Ernst & Young Entrepreneur of the Year Finalist and frequent best practices judge, William is a former Fortune 50 technology executive and database engineer. He provides Enterprise clients with action plans, architectures, strategies, and technology tool selection to manage information.

William McKnight is an Analyst for GigaOm Research who takes corporate information and turns it into a bottom-line producing asset. He's worked with companies like Dong Energy, France Telecom, Pfizer, Samba Bank, ScotiaBank, Teva Pharmaceuticals and Verizon — Many of the Global 2000 — and many others. William focuses on delivering business value and solving business problems utilizing proven, streamlined approaches in information management.

He is a frequent international keynote speaker and trainer. William has taught at Santa Clara University, UC-Berkeley and UC-Santa Cruz.

11. About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.

12. Copyright