Atos

Maven Wave™
an atos company

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

Table of Contents

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

It seems like the business world has gone crazy for machine learning (ML) and artificial intelligence (AI), viewing the technologies as the essential keys to the very future of the enterprise. In fact, a recent survey from ESIThoughtLab found that two-thirds of business leaders see AI as critically important for their future. And yet, the respondents report that only 25% of AI projects are in widespread deployment and 40% of all projects are generating negative or no returns. Further, the average ROI for AI programs is a reported 1.3%.[1] How can technology be so critical to success and, at the same time, produce such underwhelming results? Something is clearly wrong with this picture.

The missing answer to the riddle of high expectations coupled with low returns can be found in a gap between the aspirational, future state, and the hard realities that exist in the here and now. AI and ML have been a part of the conversation for some time but the reality is that the development of these capabilities is still very much at the beginning stages. In order to achieve the potential of ML and AI, fundamental factors need to be addressed and this includes the need for a comprehensive data science modernization effort. The power and potential of ML and AI won't be fully realized until the proper foundation is in place and all of the necessary ingredients to success are present.

Fortunately, a confluence of new technologies centered around the public cloud have created the perfect conditions for the widespread adoption of data science modernization, and many examples exist that can help show the way forward to success. Organizations can recognize the potential impact of AI/ML solutions and the challenges that lie ahead for adoption by first considering the difficulties associated with traditional, on-premises solutions. These aging solutions have limitations that prevent enterprises from addressing these changes. Understanding the shortcomings of traditional solutions sheds light on the ways new solutions can succeed.

In this white paper, we take a look at:

- the components that comprise a successful data science modernization effort
- how existing on-premises solutions illustrate both the challenges and opportunities that data science modernization addresses
- a real-world roadmap to beginning the journey
- discuss some of the challenges that are likely to emerge along the way

Our goal is to show how exceptional and sustainable results can arise from a comprehensive data science modernization effort at a leading, 21st-century enterprise.

The Basics of Data Science Modernization

When it comes to AI and ML, half-baked ideas and simplistic explanations often dominate the conversation and wind up standing in the way of deep comprehension and actual results. Therefore, it helps to start with understanding all of the components that comprise data science modernization so there's a solid foundation to support the examination of AI and ML.

Specifically, cloud, data, analytics, discipline, and security are essential components to any data science modernization effort.

[1] ESIThoughtLab, "Driving ROI Through AI", September 2020

Cloud: It's sometimes hard to remember that the adoption and growth of the cloud was a controversial and hotly debated topic just a few short years ago. That is no longer the case, as it is now accepted wisdom that the cloud offers flexibility, scalability, and potential for cost savings, all of which are truly transformative. This is certainly the case for data science modernization efforts as the cloud is integral to developing new capabilities that are universally distributed, readily scalable, and capable of providing the flexibility that is essential for innovation.

Data: The two big challenges with data are breaking down the silos of existing sources and dealing with the multiple forms and sheer volume of new, non-traditional data. In the first case, data has often grown in an unplanned and haphazard manner with minimal oversight. Existing data needs to be rationalized so that it's more efficient. In the second case, new forms of data might include IoT data, voice, video, or countless other sources. Provisions must account for the nuances of non-traditional data.

Analytics: AI and ML lie at the heart of advanced analytics capabilities that deliver enhanced insights and, in turn, drive an improved user experience and increased cadence for innovation. The tools for doing so are readily available, but efficient and effective utilization of the tools requires thoughtful planning. This will create a program atmosphere that harnesses the tools in a way that delivers sustained results.

Discipline: New development practices are essential for taking full advantage of the positive attributes that accrue from the cloud, including data and analytics, to deliver on the promise of accelerated insights. However, arriving at an elevated state of continuous integration/continuous delivery (CI/CD) is no easy task. After all, it touches many areas of the enterprise and requires a reimagining and realigning of people, processes, and responsibilities to deliver sustained gains.

Governance: Historically, traditional data solutions often grew in an uncoordinated manner, and the resulting "system" was usually unified and secure in name only. Too many band-aids and workarounds arise as a result, leaving data vulnerable to intrusion and compromise. A data science modernization effort has the effect of uniting governance and maintenance comprehensively and, as such, offers the opportunity to dramatically improve data security.

Taken together, these elements comprise a fulsome alternative to the many shortcomings of existing regimes. By making the most of current technologies and practices, a data science modernization effort can correct for past deficiencies and set the stage for the innovation-led growth that characterizes a healthy enterprise. However, the task is not an easy one and should roll out over an appropriate period of time using a thoughtful strategy. Taking a look at the many shortcomings in existing, on-premises solutions offers a good view of the current challenges as well as the many ways in which improvements can be made.

Limitations to Traditional, On-Premises Solutions

Examining existing, traditional methods can lend insight into how new delivery methods and practices are so differentiated. Getting an understanding of the limitations and drawbacks of the current state is a helpful way to look at what is possible with a modern approach. One method is to look at the example of SAS, the software suite that has long dominated the data analytics space.

SAS revolutionized the field of data analytics when it was introduced in 1976, and it grew into the largest privately-held software company in the world. SAS rode the initial wave of corporate data insight and, in recent years, has attempted to adjust to the new paradigm of cloud-based platforms and solutions. However, the ascendance of the cloud, and all of the advantages that it brings, highlights just how much the new model is a challenge to the traditional, installed, on-premises form of service and delivery.

Talent: The advent of R followed by Python, TensorFlow, and other data science compute platforms has led to a "changing of the guard" movement among data scientists and analysts. Workers with the skills required to run older platforms, such as SAS, are disappearing fast, and they won't be replaced as those languages are no longer widely taught. In order to secure the talent needed to effectively navigate data science modernization, organizations must accommodate available skillsets.

Portability: A major drawback of traditional approaches is the "carved in stone" relationship between data, software, and process. While this made sense 20 or 30 years ago when technology was inflexible and development processes were strictly waterfall, the current demands of business and how development is undertaken have long ago rendered past practices obsolete. Traditional, on-premises, proprietary solutions are far too constricting. For the present as well as the future, flexibility that allows for maximum portability will be necessary.

Scalability: It is extremely difficult to quickly and effectively manage scale with legacy solutions, and doing so often incurs large, additional costs for new licenses. One of the most important facets of cloud computing is the ability to scale up and scale down as demands warrant. For development, this allows for a greater degree of flexibility to innovate. At the same time, scalability in production provides the opportunity to meet needs as an organization grows, manage costs, and quickly respond to new opportunities when they present themselves. Existing solutions simply can't offer this degree of flexibility.

Performance: The one good thing about legacy systems is that their performance is largely reliable and known. However, the ability to increase performance is limited, and there is little or no wiggle room to innovate or respond to new technologies such as ML, AI, or deep learning. As with scalability, these systems are simply too rigid to meet the demands of the modern enterprise. Cloud-based solutions offer a virtually unlimited number of options perfectly suited to both the foreseen needs as well as emerging opportunities.

Maintainability: Like an old car that requires custom parts, legacy systems are facing a future where maintainability will become more and more difficult. As described above, talent is already harder to come by, and the static, waterfall-generated systems and processes are equally as unwieldy as users are locked in. New solutions, on the other hand, can take maximum advantage of open source inputs, and the use of modular design means the resulting systems are more easily modified and improved.

Cost: Existing solutions like SAS were well worth the price when they were the only game in town, but that day has long passed. It is not hard to imagine a time in the very near future where legacy solutions start to face an accelerating cost curve that will render them woefully expensive to maintain. Furthermore, there can be hidden costs that complicate TCO, repeating costs, and costs of unused resources that must be purchased due to an inflexible pricing structure of traditional, on-premises solutions. Cloud-based, dynamic alternatives have dropped dramatically in terms of price over the past few years, and there is no end in sight to this trend. Taken together, the rising cost of the old compared to the improving price structure of the new is too practical to pass up.

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

To summarize, it's not a pretty picture for legacy environments, and organizations must act quickly to address the situation. Fortunately, options exist to begin the journey and tackle the challenges and opportunities that await.

Data Science Modernization in the Real World: A Roadmap

Theoretical expositions on the wonders of new technology don't carry much weight in the real world, and for good reason. The latest and greatest tools often look very good on paper, but getting it to produce results in the face of an existing enterprise is another thing entirely. With that in mind, it is important to understand the best way to start the migration journey to new and better processes and systems.

The longest journey begins with a single step, which is true of the journey to data science modernization. Existing systems and processes may be decades old and deeply entrenched in the fabric of the enterprise, making it next to impossible to make a wholesale transformation over a short period of time. With that in mind, the best course of action is to embrace a path that begins with a workshop, develops a POC, initiates a pilot, and ultimately spawns multiple projects to drive data science modernization widely across the organization.

Workshop: The first step is to build the case for moving your data science operation. This step begins with taking a holistic view of the current state of the system including current use cases, cost, and technical requirements. Organizations should evaluate use cases based on domain and complexity and examine cost based on factors such as licenses and support. This way, there will be a complete understanding of needs and requirements. Finally, in what is typically a two-day workshop, Python and R training kicks off for legacy staff that may only know SAS or other legacy systems.

POC: Once there is a solid understanding of the current state, the next step is to identify and move a handful of cases that are well defined and capable of delivering meaningful returns when completed. The data migration process begins for one or two of these cases as well as an exploration using Python / R. The goal is for Maven Wave to "teach the team to fish" as the collaboration delivers discovery results. (Note: It's possible to skip the POC stage and go straight to Pilot if there is a high level of confidence in the proposed solution.)

Pilot: As a next step, the stakes are raised as efforts broaden. With a pilot, all new models and use cases execute in the cloud, and templates deploy, making it simple to expand and scale all programs. The goal is to build from strength to strength in an organic manner that is specifically aligned with business needs and objectives. At this stage, the one-offs of POCs are now starting to intertwine and produce dynamic, exponential growth and returns.

Projects: The project stage pulls the pieces together as all existing and new use cases move to the new framework. The data science modernization effort starts to mature as multi-tenant hosting, visualization, and dashboards take shape. Data strategy evolves to encompass all distributed data pipelines (e.g., Airflow, Spark) as well as scalable databases (e.g., Snowflake, BigQuery) and governance for users and groups is fleshed out. In all, the goal is to evolve to the point where a mature model lifecycle is achieved.

Production: An example of production is prepping models for this space in Vertex AI. Deployments are handled outside of the platform using a combination of deep learning VMs made by Google, IAC In the form of Terraform, containerization via Kubernetes, API hosting via Apigee and/or pubsub, and some overarching CI/CD.

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

The development and execution of these steps are often difficult for business leaders to conceptualize and, therefore, difficult to pull the trigger on. However, a staged approach that begins with small steps while also being mindful of the big picture is more readily achievable than may be expected. Importantly, it gets the ball rolling for what is most likely a make-or-break scenario for many companies. It's time to get started.

> The best time to plant a tree was 20 years ago.
> The second best time is right now.

– Chinese Proverb

Challenges to Implementation of a Data Science Modernization Effort

Any foundational change is going to face challenges, and this certainly applies to data center modernization. If it was easy, everyone would be doing it.

Some of the challenges that often arise include:

1. Slow ROI: It is an objective fact that ROI for data science modernization is going to start low and grow slowly over time. In fact, the ESIThoughtLab survey quoted at the beginning of this paper reported that firms are only seeing an average of 1.3% return on AI investments. If that's the case, then what are the motivating factors for adopting data science modernization? The survey has answers: AI leaders enjoy a 5x advantage compared to followers and also report a 3x increase in revenue, efficiencies, and lower costs.

In other words, results may be low at first, but followers can expect to find a landscape where their more aggressive peers enjoy accelerating gains that could leave them in the dust.

2. Ineffective project management: New solutions require new ways of thinking, and the existing team and processes may not have the requisite skills to deliver effective change. That's why it pays off to assess current capabilities and fill gaps where necessary.

Often, the best choice is to find an outside resource that can help with the lift to get to a higher level of performance without having to add excessive headcount cost and ongoing operational costs. In this scenario, outside experts should ensure the enterprise team "knows how to fish" and not only run but also grow the business once the experts make their handoff.

3. Insufficient investment: Finding the proper level of investment is as much an art as it is a science, but it is far more common to find a situation where investment is insufficient rather than excessive. A successful program of workshop/POC/pilots can start with a lower level of investment, but it pays to adopt a frank and honest approach to accurately gauge the amount and types of investment required to deliver sustained and meaningful results.

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

4. Regulatory limitations: As is so often the case with regulations, those that apply to AI and ML tend to lag behind the development of new capabilities. Care should be taken to make sure that new development is in line with current regulations and that policies and procedures adhere to industry best practices. It is wise to be prudent in an area where there are still gray areas to be explored.

5. Ethical risks: Many examples of unintentional bias have been identified in the use of AI and ML, from facial recognition programs to hiring algorithms. The stakes can be a matter of life or death, particularly in matters relating to healthcare. There should be steps taken to ensure the reliability of core data and processes and checks in place to check for unintended bias.

6. Data integrity and security: It's unquestionable that data security is essential in this era of ransomware and rising regulatory requirements. It's also critical to understand that the quality of data is directly linked to the quality of outcomes of AI/ML operations. For these reasons and more, a program that examines data integrity and security must be a first-line concern rather than a defensive operation. The highest standards will likely lead to a higher quality of end results — as well as increased innovation. Put another way, taking care of data issues now will likely pay exponential dividends down the road.

7. Challenges that outmatch capabilities: As with most technologies, there can be a tendency to deploy a solution "hammer" to a problem "nail." The fact of the matter is that there are many different alternatives available in an AI/ML toolbox and the task at hand should match the means selected to solve the problem. In most cases, the simplest solution will wind up being the best solution.

8. People: When it comes to talent, this is a "seller's market" with far more openings than there are qualified candidates. Exacerbating the problem is a "more is more" knee-flex reaction that measures AI/ML program viability in terms of the number of data scientists deployed.

Addressing the current shortage of talent and laying the groundwork for the AI/ML-centric enterprise of the future should include:

- Focusing on the proper fit of talent to the task at hand
- Developing internal training capabilities
- Building systems that democratize data science by putting intuitive tools in the hands of qualified personnel, i.e. 'citizen data scientists".

9. Institutional buy-in: All of the previously mentioned points are important, but they can easily be for naught if institutional buy-in is not active and widespread. Small, business-led efforts might show some minor results; new insights can accrue from data scientists; tech efforts at ad-hoc data initiatives might produce some efficiencies but wholesale and sustainable success can only be achieved with a commitment that is forcefully and consistently repeated from the top of the enterprise.

The Inevitability of Data Science Modernization During the Machine Learning and AI Revolution

Data Science Modernization: A Prerequisite for Future Growth

Amazon has been one of the greatest business success stories of the 21st century, but a core component of its success is often overlooked. Specifically, Amazon CEO Jeff Bezos reportedly issued a short, seven bullet memo in 2002, which mandated that, without exception, all data and functionality had to be available and shared through interfaces going forward. Further, all of the interfaces had to be externalizable. (The sixth and seventh bullets said, respectively, that “anyone who doesn’t do this will be fired” and “thank you; have a nice day!”)

In retrospect, Bezos clearly recognized the benefits of an open data strategy, and he was able to express this mandate in seven succinct bullets. There was no mandate on what technology or methods were to be used, only that data had to be accessible through an interface, or, in modern parlance, an API. It’s an interesting contrast to the complicated jumble of data that most companies currently confront. Just as Bezos was able to mandate an “API culture” at Amazon, today’s enterprises must embrace data science modernization. The result is increased insights, enhanced cadence, and improved innovation that they not only impart but, quite frankly, also to keep up with the competition. The cold, hard truth is that “data gravity” increases as efforts expand and success builds upon success and data is paired with other data. The democratization of AI helps to create “citizen data scientists” and introduces further exponential impacts. All in all, the stakes are high, and the time is now to get started.

Companies in the midst of digital transformation too often don’t see a single business win until the end of a long technological overhaul. Maven Wave’s progressive approach to modernizing data delivers value in the earliest stages of the project lifecycle. Rather than upending everything at once, we use an agile approach to deliver the data and insights in a prioritized fashion. This approach helps justify a shift to modern platforms and keeps the business engaged and invested in a transformational change.

Additionally, Maven Wave can help you to safely, quickly, and more effectively move away from the SAS platform in favor of more modern technologies. Within 4–6 weeks, we can help build a comprehensive data platform that leverages the data and machine learning innovation of Google, while scaling quickly and easily on best in-class infrastructure.

Learn more about our SAS Modernization solution here.