Datameer

Solution Brief

Agile Data Pipelines for Cloud Machine Learning

AI and Machine Learning have the ability to transform the world as we know it. Many Fortune 500 CEOs feel the same—81% of respondents in a 2017 survey said that AI and Machine Learning were a „critical area of investment."

However, despite the hype around data science, artificial intelligence, and machine learning, much of the work is still cloistered away on disjointed data science teams. Every company wants to use it, but few know how. In fact, a report coming out of MIT Sloan showed that while 85% believed AI would give them a competitive advantage, only 20% of the respondents were actually using it.

Why is AI/ML so hard to implement? Much of the challenge comes down to the manual aspects of the machine learning analytic cycle – accessing data, preparing data, exploration and feature engineering, model validation and finally operationalization. Even with the most powerful cloud computing and machine learning algorithms, the heart of the matter lies with the data. Therefore, in order to shorten analytic cycles and gain the competitive advantage, organizations need to find ways to streamline and simplify the data processes for machine learning.

Many industries are taking their computing to the cloud to help solve these problems. What used to require expensive on-premise servers can now be outsourced in a flexible, efficient way. The scalability and elasticity of the cloud allow teams to spin up new instances in a moment's notice. This in turn gives teams the ability to iterate and optimize models faster than ever before, rapidly increasing their speed-to-insight.

But despite all these advances, many organizations still depend on highly manual processes. The fragmented workflow of data management, model validation, and operationalization continues to leave gaps in machine learning analytic cycles.

The Big Challenge for AI/ML

Today, most of the AI/ML data cycle work is done manually on an ad-hoc basis. While this initially can save time and bootstrap early insights, it leads to brittle workflows that don‘t scale. Many analytics teams depend on one or more of these manual processes:

- Accessing data
- Scripting and hand-coding data preparation
- Modeling
- Validation
- Operationalization

All the while, this disjointed process leaves key stakeholders out of touch and out of the loop. Without a continuous link between data, modeling, validation, and operationalization, the AI/ML data cycle is bound to stay inefficient and prone to errors.

Let’s look at each of these obstacles more closely.

Five Major Obstacles

Accessing the Right Data

Organizations are swamped with data. This presents both an opportunity and a challenge. The opportunity is the use of more data means better models and more accurate results. The challenge is getting to the data. Many organizations silo their data in disjointed places that are hard to navigate. Data scientists then need to make data requests from IT which can add weeks or months to their analytic cycles.

Analysts can provide more value when they’re able to access the right data at the right time. This turns the data scientist’s job into finding the proverbial needle in the data haystack. Often times they may not know what data is useful. Without a full understanding and application of the right data, organizations end up with an inefficient, or worse, incorrect model.

Reducing Time Spent on Data Preparation

Data scientists can spend up to 80% of their time on data preparation alone, according to a report by CrowdFlower. In today‘s fast-paced marketplace, this is unacceptable. When data prep takes up the majority of an analyst‘s work day, they have less time to spend on

the actual tasks of building models and drawing insights. Again, this phenomenon arises mainly from the abundance of manual data processes surrounding the organization.

While Python and R have built-in coding functions to help with data preparation, this is still a fragile process. Analysts must write custom scripts for each data preparation job, which leaves room for redundancy and human error. Even once a script is perfected, the logic remains siloed in that data model and cannot be reused elsewhere.

Choosing the Right Data

Data exploration is critical to the machine learning process. It is essential to identify which datasets are most likely to be relevant to the problem being solved. Data exploration, as well as algorithmic exploration, helps feed the feature engineering part of the data science process enabling the data scientist to more efficiently see what variables are most important to the problem at hand. Therefore, analysts end up spending inordinate amounts of time exploring and trying to understand the data. Without a modern exploration platform, this can be a very resource-intensive process.

Notebooks, Python, and R have built-in charting functions, but this is a highly manual process just to get simple histograms and distributions of the data. This method carries with it many of the flaws noted above. In addition, it does not offer any ability to do multi-dimensional analysis and relationship analysis, both of which are essential to understanding the true meaning of the data. When analysts move beyond surface-level exploration, they can see which variables and features are most relevant.

Validating and Iterating

Model validation is also a highly manual and difficult process today. Data scientists will explore model output in Excel, or the simple charting functions mentioned above can be applied to model testing output, but they only provide simple, one-dimensional views of the results.

Data scientists need to dig deep into the testing data to explore the results of a model in any number of dimensions and down any number of exploration paths in order to truly see how a model behaves. This process should also involve the business stakeholders, so they can see how the model behaves and add feedback from a business perspective. But when this process is separated from the stakeholders and the rest of the organization at large, it becomes hard for analytic teams to get the feedback they need.

Operationalizing Models

In the traditional AI/ML data cycle, data models must undergo a time-consuming "re-implementation" process after validation. The data scientist throws the model over the wall to IT, who then codify and build a custom data framework to feed the model. This is a lot of duplicated work, not to mention highly error-prone. At this stage, teams must also take into consideration potential security, governance, and compliance requirements that need to be applied to the machine learning problem. Organizations need to be accountable to stakeholders about where the data goes and what happens to it—which adds another level of complexity.

With all these problems and resource-intensive processes, it is no wonder why machine learning analytic development and deployment cycles are so long. Sometimes they never even see the light of day within the business.

Modern Data Pipelines to the Rescue

Modern data preparation, exploration, and pipelining platforms such as Datameer provide the proper data foundation and framework to speed and simplify machine learning analytic cycles. They provide the self-service tools for preparation and exploration, scale, automation, security and governance to alleviate all of the aforementioned gaps in the machine learning analytic process. Let's look at how.

Access More Data

One of the most important steps in the machine learning process is making sure your team has governed, self-service access to more data, so they can find the right data for the problem they are solving. An enterprise data preparation platform brings together data engineers, who manage access to data, and data scientists, who need access to the data and easy preparation steps, onto a common platform to collaborate. This helps organizations curate a large number of datasets that can be cataloged and available for analytic purposes.

The more data that is available and curated for use, the more assets an analyst can:

- Explore to uncover the relevance and meaning,
- Prepare and engineer to get ready for the machine learning process,
- Feed their machine learning models to train and test their models.

The platform's built-in catalog gives the data scientist the ability to understand and find more data related to their questions. This also means that data scientists can access and explore the data in one place instead of pulling it from various sources. In the case that

the analyst needs access to a different dataset, the data engineer can fulfill this request in a matter of minutes.

Prepare Your Data

We discussed earlier the highly manual data preparation processes that often occur in data science analytic cycles. This dramatically lengthens the cycle and hinders efforts for re-using logic and curated datasets.

A modern data preparation platform such as Datameer Spectrum dramatically reduces time to insight by providing self-service interactive interfaces for teams to prepare data – clean, shape, blend, aggregate, organize – as well as explore the data at scale. Spectrum offers a spreadsheet-style interface with a robust suite of over 270 functions, many of which are Excel-like, making the approach very familiar to a data scientist. This non-technical interface makes it faster and easier to prepare the data for machine learning purposes.

Explore and Engineer

A major part of the data engineering component of the data science process is to dig in and understand how the various datasets, and the data as a whole, relates to specific problem the data scientist is trying to solve. This involves two key aspects: data exploration and feature engineering.

Datameer Spectrum goes beyond your typical data preparation tools by offering unconstrained visual data exploration at scale. This allows a data scientist to interactively explore even the largest datasets to identify relevant aspects and patterns of the data to see the relevance to their problem.

Visual Explorer in Datameer Spectrum allows interactive exploration of datasets into the many billions of records with response times in the seconds. It uses a unique dynamic indexing technology that allows a data scientist to explore across any dimension, value and metric in the data, and switch at any time. This enables a rapid, “fail fast” exploration of any number of paths within the data to see the most important nuances.

Spectrum also offers a number of unique features for algorithmically exploring different attributes in the data. Data relationships, decision trees, clustering, mathematical and other built in algorithms and functions allow the data scientist to do a first pass at feature engineering to see what attributes may be most relevant to the problem.

Validate Your Models

Analytics is an agile, iterative process—not a linear one. Data is prepared and models are created, tested and validated multiple times in a cycle to try and get the best possible model. This ability to iterate on model feedback and validation is a crucial piece of this process. By feeding modeling test results back into the visual exploration tools, the analyst can speed their model validation process.

Datameer Spectrum can ingest model test results, combine the results with the test data itself, and allow the data scientist use Visual Explorer to examine every aspect of the model behavior interactively. They can compare results and look at the effect of different variables on the model when expanded in various directions and dimensions.

In addition, because Visual Explorer is easy and graphical, the data scientist can sit down with the business stakeholders and show them how the model behaves. This brings the business team into the process, enabling them to give business validation of the model and further shrinks the overall time to insight.

Operationalize and Automate

The re-implementation, or operationalization, aspect of data science has long plagued its' ability to deliver value to the business. The complex, custom coding process can easily add weeks or months between the time a model is deemed ready, and when the business can use it. And as models evolve, the delivery of new models get delayed as well.

What data science needs is a scalable, automated data pipeline platform where data scientist can collaborate with IT and data engineers to simply plug-in models to existing secure, governed data flows. Datameer tackles this problem directly.

Once the data preparation has been performed, this data flow can be operationalized and automated to run at scale. Once the model is complete and validated, Datameer Spectrum's plug-in architecture allows models to be plugged into these data flows to produce machine learning enriched output.

This not only speeds the operationalization process, but keeps all the stakeholders happy:

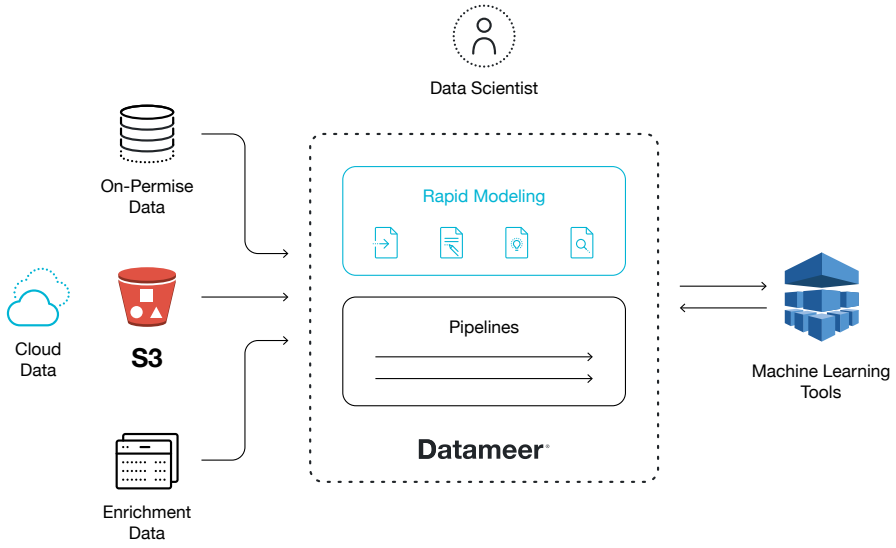
- IT teams can use the plug-in architecture and automation tools to productionalize the ML-enriched data flow easily
- Data engineers can properly secure and govern the use of data for the ML-enriched data pipelines

- Data scientists can see their hard work pay off quickly and efficiently for the business teams, as well as get a continuous stream of information on how the model behaves in real-life
- Business teams get their data science models faster to create real business value

Once the model has been trained and tested, analysts can set up automated data pipelines that process new data, feed it into the model, and then produce results which are used by the downstream business stakeholders. This becomes a continuous process with minimal intervention from the data scientist and maximal results. On top of it all, each step in the process is logged and monitored for security and governance purposes.

Taking Advantage of the Cloud

All these features come together when integrated with industry-standard cloud services. Being „cloud-native“ is more than just a buzzword—it‘s a paradigm that provides a seamless experience for everyone involved.



Faster AI/ML Analytic cycles

Operationalize AI/ML Models at Scale

Easier Feature Engineering

Deep Security and Governance

To be cloud-native, Datameer Spectrum on Amazon Web Services (AWS) leverages key AWS services including:

1. **Use of S3 Cloud Object Storage** – Spectrum stores and manages data on the S3 object storage. This provides persistent storage services and enables Spectrum to separate compute and storage services it uses, enabling these to be scaled independently.

2. **Elastic compute with EMR clusters** – Spectrum uses EMR for compute capacity. EMR clusters can be scaled independently by customers allowing Datameer to take advantage of burst computing capacity.

3. **Integration with AWS IAM** – In Spectrum, security is integrated the AWS Identity and Access Management system.

4. **Integration with AWS KMS** – for encryption key management, Spectrum integrates with the AWS Key Management System.

5. **Integration with AWS Management Console** – Spectrum instances and the underlying EMR and S3 services Datameer uses, can all be managed within the AWS Management Console.

Data platforms that integrate with external data sources, cloud services, and security layers provide an all-in-one solution for today's analytic needs.

Conclusion

Where are the bottlenecks in your agile data organization? Even if you're using modern machine learning on the cloud, you're missing an opportunity if you're still using manual processes to access prepare and explore your data, as well as validate models. With the agility and scalability of a modern data platform, analytics teams can make better decisions and provide faster results.

About Datameer

Datameer provides hybrid-cloud analytics data management products that help organizations have successful, lower risk cloud journeys. Datameer Spectrum enables organizations to create secure, governed and scalable data pipelines and DataOps for easier, lower risk migration of analytics workloads to the cloud with reduced data engineering and cloud infrastructure costs. Datameer is a trusted platform at leading enterprises globally, including Citibank, Royal Bank of Canada, British Telecom, Aetna, Optum, National Instruments, Vivint and more. To learn more, please visit www.datameer.com.