WHITE PAPER

Advance Your Analytics

Commentators suggest that machine learning will soon be ubiquitous and the basis of competitive advantage in many industries.[1] As a consequence, investment in machine learning and artificial intelligence technologies has increased rapidly and is predicted to grow even stronger.[5] Despite this investment, however, Brynjolfsson et. al. note a "modern productivity paradox" and conclude that "implementation lags" have so far prevented machine learning and artificial intelligence from realizing their full potential.[2] Put simply, the expertise required to successfully deploy machine learning and artificial intelligence at-scale in large organizations is not evenly distributed—and while many organizations are capable of delivering proof-of-concept solutions, relatively few have demonstrated they can deploy large numbers of production solutions at scale.

Table of Contents

Machine learning and artificial intelligence are first and-foremost a data problem. The importance of standardizing the structure and processing of analytic datasets has long been recognized,[3] but progress in this area has been hindered by a proliferation of tools, technologies, data silos, and "one-pipeline-per-machine-learning-process" thinking. The consequences of the current approach are: poor data scientist productivity, with data scientists spending between 50 and 80 percent of their time wrangling data rather than building predictive models;[4] time-to-market issues; and failure rates for analytic initiatives estimated by Gartner as being in excess of 80 percent.[5]

The advanced analytics process can be thought of as consisting of three main components.

1. **Feature Engineering**: this is a data management, integration, and manipulation task.
2. **Model Training**: also known as model creation or simply modeling. This task involves taking existing features and designing new features that can then create a mathematical model that can, for example, predict future values with sufficient accuracy to help inform business decisions.
3. **Deployment**: the final task is to take the features from part one, the trained model from part two, and then apply a scoring function to production data to generate a prediction.

The basic tenets of Teradata's Analytics Strategy are that (a) organizations can only successfully scale their machine learning and AI initiatives if they pay greater attention to feature reuse and model deployment and that (b) feature engineering and model scoring are directly aligned with Teradata's core value propositions—and as such are a perfect fit for Teradata Vantage.

Data science pipelines should be reengineered to populate and maintain an Enterprise Feature Store,

[1] Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence*; Harvard Business Review Press.

[2] Brynjolfsson E, Rock D, Syverson C, (2017) *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*; The National Bureau of Economic Research.

[3] Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59 (i10).

[4] Dasu, T, & Johnson, T (2003) *Exploratory Data Mining and Data Cleaning* (Vol. 479). John Wiley & Sons.

[5] White A (2019), *Gartner Predicts 2019: Data and Analytics Strategy*; https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/. Retrieved 2020-21-04.

materialized as tables in an Analytic RDBMS, so these features can be reused to both train and score multiple different models. These curated collections of variables with proven predictive value are already dramatically improving data scientist productivity and time-to-value for new analytics in leading organizations.[9, 10] Deploying these "feature stores" on their existing Teradata systems will enable Teradata customers to avoid data movement and duplication, reducing both TCO and latency, while also leveraging Teradata Vantage's industry-leading performance and scalability for the manipulation and processing of large analytic datasets.

The model training task is typically completed using carefully chosen samples of historic data. By contrast, the model scoring process often requires access to complete and up-to-date analytic datasets in the feature store. Additionally, the model scoring task is typically (a) mission critical, (b) requires predictions to be made available at an operational endpoint, and (c) is increasingly being executed in near real-time. Teradata systems are highly available and provide industry-leading mixed-workload management. They are also typically directly connected to operational endpoints across multiple channels and, crucially, support the "tactical" queries that characterize near real-time model scoring with response times measured in tens of milliseconds. Teradata customers can use the most appropriate tool to train any particular predictive model including the Teradata and/or in-database model training functions where these are appropriate, while also ensuring they can score models at scale directly against production data. This is achieved by using subsets of data from the feature store to train a predictive model either in-database or externally. Where the model has been trained externally, it is then imported into Teradata Vantage through a process known as "Bring Your Own Model" (BYOM) and is used to score production data in the feature store. This process, and these technology capabilities, make Teradata Vantage the ideal platform for the operationalization of machine learning in large and complex organizations.

Problem Statement

Data scientists and engineers have established a way of working known as pipelines, which are typically end-to-end processes designed and built to solve a problem on a project-by-project basis. The pipeline starts with code for feature engineering (also known as data wrangling). For small scale developments, testing, and "science projects," this is a good approach, as it is both efficient and ensures repeatability. Ensuring the same result can be consistently and reliably reproduced whenever an experiment or an analysis is undertaken is a key consideration in research and, increasingly, in regulated industries. Repeatability is usually achieved by storing the machine learning pipeline as code in a versioning repository, like git or svn. However, scaling this approach up to an enterprise level very quickly

Analytics 1-2-3

Feature Store

- Data Preparation
- Data Integration
- Feature Engineering
- Update and Reuse

Model Training

- Open to all Tools
- Production Data
- Train and Evaluate
- Export Models

Production

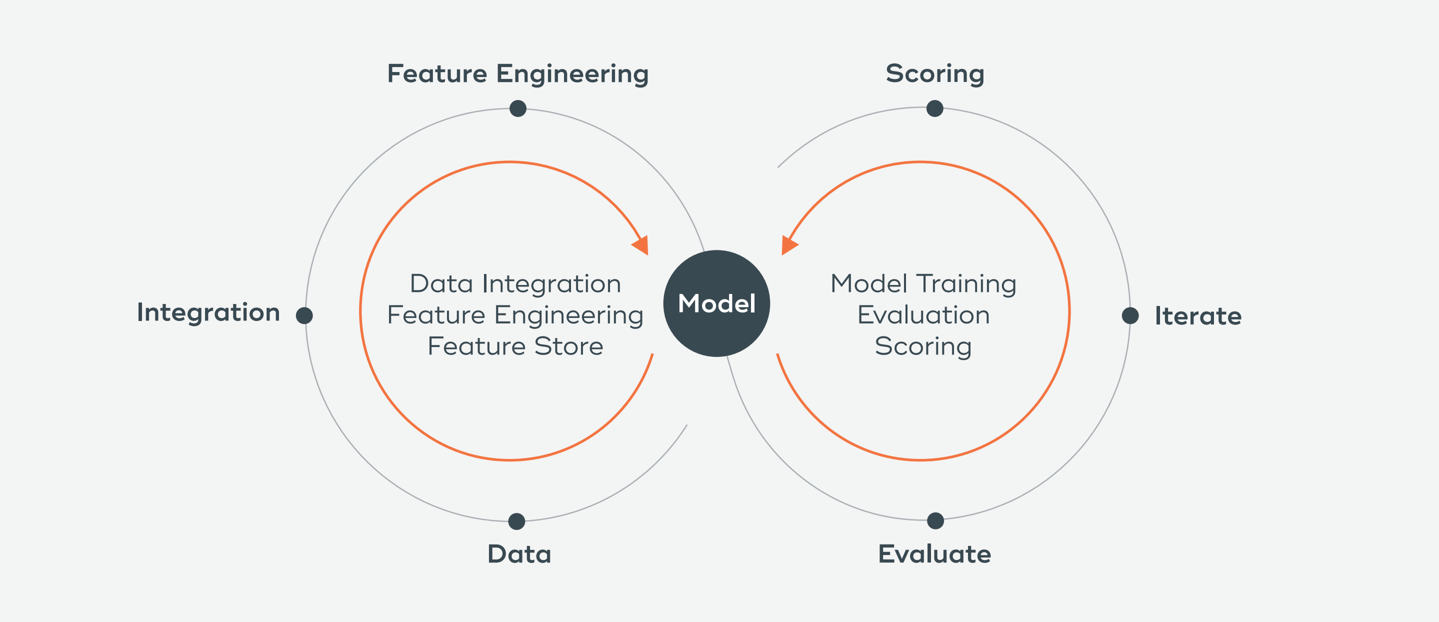
- BYOM
- Model Scoring
- Enterprise-wide
- Automation

leads to a highly inefficient process that creates silos of data and processes that are stored as code—which are often indecipherable except to the original author. Google describes the outcome of this method of working as "pipeline jungles"[6] and note that data dependencies are one of the key contributors to technical debt in machine learning systems.[7] It is also noted that data dependencies have a higher cost than code dependencies.

Data scientists are perceived as curious and inquisitive, constantly researching new tools and processes while expanding their knowledge to encompass the very latest skills and techniques. However, in practice, the data science community is very diverse and different groups tend to stick to the skills they have, focusing on becoming experts in their field. This is particularly true as it applies to analytical languages. The Python coder remains exceptionally loyal to the Pythonic method of programming, the R user will work almost exclusively with the library of scripts they have developed over time, and robust and proven tools like SAS are often highly-valued by users working in heavily regulated industries. The current reality is that no single analytics tool, language, or framework has been able to establish a dominant position in the marketplace, and for as many analytics professionals that swear by R or SAS, there are an equal number of devotees to Python. This partisan approach means that introducing different coding languages and methods to an established data science team will often be met with opposition. It is also the case that there is no objectively "best" technology for the wide variety of model training activities undertaken in large and diverse organizations, and that in very many cases a good outcome can be achieved using various different libraries, methods, and languages. In addition, data scientists will often be more efficient, comfortable, and creative using tools they are familiar with, and as such it is often not desirable to restrict model training activities to a single technology. Furthermore, the current enthusiasm for machine learning and artificial intelligence has led to a proliferation of new analytic tools, languages, and frameworks. As the market matures, it is likely that some of these approaches will fall from favor, that others will become confined to one particular niche—and that others will be consolidated into broader offerings that can secure significant market share. Predicting the outcome of this process is fraught with risk and uncertainty.

[6] Sculley D et al. (2015) *Hidden Technical Debt in machine learning systems*; Google inc.

[7] Sculley D et al. (2014) *Machine Learning: The High-Interest Credit Card of Technical Debt*; Google inc.



Finally, research shows a large proportion of advanced analytics projects never reach production and never go on to generate value to an enterprise.[5] Production means different things to different personas, but to the business production means the results of models are not only available to business users but are trusted and used on a regular basis to make decisions such as next best offer, churn reduction strategies, or retail price changes, etc. Any production system involving advanced analytic models must be scalable, performant, robust, easy-to-maintain, and secure. In particular, it is increasingly important, again especially in regulated industries, that organizations are able to understand why and how these systems made a particular prediction at any arbitrary point in the future.

Solution: Analytics 1-2-3

If organizations are currently unable to standardize on a single analytic tool, language, or framework, the question remains: how can they scale the deployment of robust, enterprise-grade analytics? It can be observed that the organizations which are most successful are those that have invested less in trying to standardize on a single analytic tool and more on getting the end-to-end analytic process right, with a particular emphasis on the activities that occupy the two ends of the analytic value chain. A flexible approach is required to quickly incorporate the newest deep learning library, or the latest analytic language and library with a sweet-spot aligned with the current task-in-hand. A solution is to decouple the different parts of the analytics process from each other and to run each component on an optimal technology. In the same way that modern compute architectures separate storage from compute, we should separate the parts of the analytics process. Decoupling the process leads to a more efficient system and we can use the principle of "polyglot programming"[8] so that appropriate tools, languages, and frameworks are applied to the tasks they are best suited to; or that tools, languages, and frameworks are chosen to reflect the expertise and experience of the user. One way of looking at the process of creating predictive models is the following 1-2-3 approach:

1. Feature engineering: this first step is for data preparation and management which still accounts for 80 percent of the cost and effort of analytic projects. If machine learning is to become ubiquitous, we need to bring very significant pressure to bear on that number. And that means that we need to stop handcrafting one data pipeline per machine learning process and to invest instead in creating curated "feature stores." These are tables of data where the columns hold transformed variables with proven predictive value that enable reuse. The feature store is a major part of the advanced analytics strategy for data driven companies such as Uber[9] who note that "Data engineering is the hardest problem in machine learning."
2. The second step in the process is model training. This is the home ground of the data scientist. It is imperative that, given the boundaries of an enterprise framework, data scientists are free to explore data and algorithms to provide robust, accurate models that will provide a solid, quantifiable ROI. This freedom should include the ability to use a variety of tools. Teradata Vantage's advanced analytics capability for data exploration and model training explicitly allows for the use of other tools and languages, where these are more appropriate. The problem here is the analytics industry has dramatically over-rotated on model creation. This is a mistake, as it is clear that production analytics at scale do not begin or end with a predictive model.[10] That means that the productivity of our data scientists is low, time-to-market of analytics applications is measured in months, and that in many cases they do not get to production. For the foreseeable future, businesses will need to use multiple different technologies to support model creation activities. Given that, the data required to train the models should come from reuse of variables stored in the feature store where possible, and any new features created should be added to the

[8] Ford N, 2006 *Polyglot programming*, http://memeagora.blogspot.com/2006/12/polyglot-programming.html. Retrieved 2020-15-04.

[9] Hermann J. Del Balso (2017) *Meet michelangelo: Uber's machine learning platform*; Uber Engineering

[10] Lin J, Ryaboy, D (2013) *Scaling Big Data Mining Infrastructure: The Twitter Experience*; Association for Computing Machinery

The fact that everything exists on the Teradata system facilitates near real-time prediction.

feature store without exception. Note that there is a discussion to be had as to whether a "model" is the result of an algorithm being trained on data or consists of the trained algorithm plus the features that created the training data. For the purposes of this paper, feature engineering and model training are treated as two completely separate activities. However, it is fully acknowledged that the iterative nature of model creation means that, in the discovery and evaluation phases, these two activities are intrinsically linked. The point being made is that once a model has been created and shown to be accurate, the feature engineering code should be migrated to the feature store update process and not left in a model-specific pipeline. Treating model creation as a separate activity allows Teradata Vantage to seamlessly incorporate models trained in external systems in addition to in-database models or models created using the Vantage MLE.

3. The third part of the process is where the trained model is deployed. Here, a scoring function is applied to the model and live production data from the feature store. This creates the predictions that are persisted on a regular basis. If parts one (the feature store) and two (the trained model) are in place, everything required exists in-database with no data movement to or from external systems required. The production process in part three is therefore simple and robust. In addition, Teradata Vantage's always-parallel-all-the-time architecture means that batch scoring workloads are both performant and scalable, meaning that new predictions can be created on live production data as often as required. In addition to operating in a batch mode, the fact that everything exists on the Teradata system facilitates near real-time prediction. Model scoring is typically an example of an "embarrassingly parallel" process and the nature of Teradata Vantage's logical, hash-based filesystem means that near real-time scoring operations in Vantage are generally "single AMP, single (logical) IO" operations that are executed in tens of milliseconds and consume very few CPU and IO resources. A completely automated system can be built around this core, including regular testing for model drift, retraining and a champion/challenger methodology for new model release into production. This is part of a process often referred to in the industry as "AnalyticOps." Teradata Vantage has a proven solution to this, but it is beyond the scope of this document.

Teradata Strengths

Teradata Vantage's unique implementation of the Massively Parallel Processing (MPP) model provides very significant advantages over competitor implementations and has given the company a flexible platform from which to continue to innovate and to lead the market. Ubiquitous machine learning implies that organizations will need to deploy between tens and hundreds of millions of predictive models in production in the near future. Teradata Vantage has already demonstrated the ability to scale machine learning vertically (by training models on more than a million observations and scoring them against more than 250M observations multiple times per day) and horizontally (by training millions of predictive models to support so-called "hyper-segmentation" use-cases and scoring them daily) in demanding production settings for some of the largest and most analytically sophisticated customers in the world.

In particular:

- By combining an "always parallel, all-the-time" architecture and processing model with a sophisticated cost-based optimizer, Teradata enables the high-performance processing of large and complex datasets that characterize many data preparation, model training, and model scoring operations.
- Because it is built on a logical, hash-based filesystem, Teradata provides O(1) access to localized data, enabling extremely high throughput and low-latency for tactical queries, like near real-time model scoring.
- By delivering an industry-leading mixed-workload management capability, Teradata Vantage enables mission-critical operational workloads (like near real-time model scoring) to co-exist with complex, resource-intensive processing (like data preparation and model training), eliminating the need to duplicate data in multiple, redundant, overlapping silos.
- By incorporating in-database processing of analytical languages like R and Python, Teradata Vantage enables data preparation, model training and model scoring activities to be performed in-database using languages and libraries that are popular with many data scientists. In particular, this can eliminate the need for models developed using external tools to be reused so they can be deployed in production and at scale.
- Tight integration with tools provided by leading analytics vendors allowing externally-trained models to be scored in production in-database. For example, the SAS scoring accelerator, enables the efficient implementation of analytic processes in production and at scale.
- Teradata's QueryGrid virtualization framework and Incremental Planning and Execution (IPE) technology enables data persisted in data lakes and across the analytic ecosystem to be transparently and performantly queried and combined with data managed in the platform's database, enabling rapid and flexible data exploration.
- Fully accountable AI is possible using a combination of temporal tables which allow data to be viewed in the exact state it was at a particular moment in time. This, coupled with full logging of database queries and model metadata and management in database, means that a precise investigation of why a particular model made a particular prediction on a certain date can be made, providing for auditability and repeatability.
- Teradata Consulting BYOM assets and frameworks enable models that have been developed in a wide range of analytic tools to be imported and then scored at scale against production using a variety of methods including PMML, SQL conversion, and native code running in-database.
- The AnalyticOps accelerator allows full automation of the analytics process, including the use of CI/CD pipelines to maintain feature stores and model management at scale. This means the time to deliver analytics is greatly reduced.
- By providing true hybrid cloud deployment capabilities, Teradata Vantage provides on-premise, virtual private cloud, and public cloud platform deployment options. As Teradata Vantage is exactly the same product offering irrespective of the underlying infrastructure platform, it eliminates the need to reengineer existing applications. This ease of portability both lowers the barriers for organizations to adopt a new platform model and provides portability across leading cloud ecosystems.

Conclusion

The twin realities of rapidly accumulating technical debt in machine learning deployments and current market dynamics mean that organizations need to take a more holistic view of the machine learning process. While we should continue to compete for model training workloads wherever this is appropriate, we need to also acknowledge there is potential to bring a large amount of new workload onto Teradata systems through customers migrating their current advanced analytics solutions to Teradata as a platform for data management and ongoing production. Capturing these parts of the machine learning process will increasingly be our focus, because migrating these workloads to Teradata Vantage has the potential to create very significant additional value for our customers by helping them to address productivity and time-to-market issues, as well as by reducing failure rates.

Teradata's traditional value proposition—scalable, enterprise-ready, high-performance processing of large and complex datasets—is as important to our customers' ability to undertake predictive and prescriptive analytics at scale in this decade as it was for their ability to scale descriptive analytics in previous decades. It will therefore be increasingly important for Teradata sales teams to understand the different parts of the advanced analytics process and to recommend that the parts most suited to Teradata's strengths are executed in-database.

Advanced analytics projects involve data integration and manipulation; data scientists call this process feature engineering and, in many projects, a large amount of resources, both machine and human, are consumed to create the features that are used to train a predictive model. A recent industry trend is to pool these generated features from multiple projects in an Enterprise Feature Store[9] that can serve many applications in production. Teradata Vantage is the perfect place to generate and host this feature store. In addition, most businesses struggle with the last mile of pushing advanced analytics into production. As with feature engineering, a large amount of workload is generated when a scoring function processes live data from the feature store record by record, applying the model and producing predictions. This scoring process can generate tens of millions of queries a day and in Teradata Vantage we have an enterprise-ready platform that is scalable and robust enough to guarantee the continued production of quality predictions.

About Teradata

Teradata is the cloud data analytics platform company, built for a hybrid multi-cloud reality, solving the world's most complex data challenges at scale. We help businesses unlock value by turning data into their greatest asset. See how at **Teradata.com**.

About the Authors

Chris Hillman is a London-based principal data scientist on the international advanced analytics team at Teradata. He has more than 20 years of experience working with analytics across many industries.

Martin Willcox is the VP of Technology for Teradata EMEA. He is jointly responsible for driving sales and consumption of Teradata solutions and services throughout Europe, the Middle East, and Africa.

17095 Via Del Campo, San Diego, CA 92127 Teradata.com