WHITE PAPER

Analytics 123: Enabling Enterprise AI at Scale

How de-coupling the pipeline can deliver ROI on machine learning and AI

By Martin Wilcox, Vice President Technology (EMEA), Teradata
and Dr. Chris Hillman, Principal Data Scientist, Teradata

09.20 / DATA ANALYTICS / WHITE PAPER

teradata.

Table of Contents

Business leaders recognize that machine learning and artificial intelligence (AI) will soon be ubiquitous and the basis of competitive advantage in their industry. McKinsey suggests that by 2030 70 per cent of businesses will have adopted at least one form of AI[1] and Gartner predicts that by 2022 90% of corporate strategies will explicitly mention analytics as an essential competency.[2] As a consequence, investment in machine learning and AI technologies has increased rapidly and is predicted to grow even more strongly. KPMG suggests investment will jump from $12.4 billion currently to nearly $150 billion by 2025.[3]

Despite these investments, hopes and expectations, many businesses are struggling to see returns from machine learning and AI projects. Sixty-five per cent of executives worldwide report that they are not yet seeing value from their AI investments.[4] A well regarded and often referenced academic paper from Brynjolfsson et. al. highlights a "modern productivity paradox" and concludes that "implementation lags" have so far prevented machine learning and artificial intelligence from realizing their full potential.[5]

1 https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

2 https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019

3 https://home.kpmg/lu/en/home/insights/2019/04/khube-mag/intelligent-automation-edition/the-vast-world-of-intelligent-automation.html

4 https://www.forbes.com/sites/gilpress/2019/10/17/ai-stats-news-65-of-companies-have-not-seen-business-gains-from-their-ai-investments/#447ffcba19f4

5 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 https://www.nber.org/papers/w24001.pdf

teradata.

The Gap Between Promise and Delivery

What lies behind this discrepancy between the promise and the delivery of AI and machine learning? The answer lies in the ability to deploy analytics at speed and scale. Machine learning and AI are first and-foremost a data problem. The importance of 'Tidy Data'[6] and standardizing the structure and processing of analytic datasets has long been recognized but progress in this area has been hindered by a proliferation of tools, technologies, data silos and "one-pipeline-per-process" thinking. Plus, the expertise required to successfully deploy Machine Learning and Artificial Intelligence at-scale in large organisations is not evenly distributed-and whilst many organisations are capable of delivering proof-of-concept solutions, deploying analytics at production scale is several orders of magnitude harder and something that few business have been able to do successfully.

The problem is rooted in inefficient, non-optimised processes that restrict organisations' ability to make best use of their investments in data assets, technology, and skillsets.

To compete today, let alone survive and become an analytics-driven Business of the Future, organisations must act now to build flexible, repeatable and accountable data processes that provide solid foundations for AI. Teradata's Analytics 123 strategy establishes a straightforward roadmap for both business and analytics leaders that creates robust, efficient and easily deployed processes that ensure machine-learning and AI projects live up to their promise and deliver real business value. Analytics 123 decouples the different elements of the analytics process and ensures appropriate weight is given to each. Stage one is feature engineering with reuse at its heart. Stage two gives data scientists the flexibility to use their preferred tools to create predictive models with value to the business. Stage three deploys those models to score live data.

What lies behind this discrepancy between the promise and the delivery of AI and machine learning? The answer lies in the ability to deploy analytics at speed and scale.

Prepare data

50-80% of time taken preparing raw data

- Data Integration
- Data access and exploration
- Data cleansing
- Feature engineering
- Feature selection

Train model

Fit ML algorithm to the training data:

- Algorithm selection
- Test and training data-set split
- Model training and evaluation
- Model optimisation
- Model export

Deploy model

Operationalize model to predict outcomes:

- Write-back new features
- Import model to model repository
- Operational scoring
- Business process integration
- Model monitoring

6 Wickham H (2014) Tidy Data, Journal of Statistical Software https://www.jstatsoft.org/article/view/v059i10

teradata.

The End of The Pipeline

Today, most organisations take a tightly integrated 'pipeline' approach to analytics projects. Pipelines are typically end-to-end processes designed and built to solve problems on a project by project basis. They start with source data and write code for feature engineering (also known as data wrangling). For small-sale testing and research/experimentation projects this approach works well. Resources are focused and used efficiently, and the approach lends itself to repeatability as the whole pipeline can be stored as code in a versioning repository (such as git or svn). The same result be consistently and reliably reproduced whenever an experiment or an analysis is undertaken.

However, scaling this approach up to an enterprise level quickly leads to inefficient processes creating silos of data and code. Individual teams will often duplicate effort, engineering almost identical features from the same data but siloed within their own pipelines and inextricably linked to the predictive models they support. This leads to poor productivity among data scientists who spend between 50 and 80% of their time wrangling data rather than building predictive models.[7] Features are created and used in isolation. Other teams working on related issues will be unaware that much of the data wrangling they need to do may have already been done. As the demand for predictive and prescriptive analytics rises this 'data wrangling overhead' is simply unstainable.

This low productivity and high cost contribute to the disappointing return on analytics investments and poor progress towards more data-driven organisations. Not only is it adding expense through duplicated effort, but extended project timelines and slow time to market erode both impact and trust in the value of machine learning across the business. Analytics in production means different things to different audiences but to the business it means that the results of models are trusted and used on a regular basis to produce value by making decisions such as next best offer, churn reduction strategies or retail price changes. Any production system involving advanced analytic models must be scalable, performant, robust, easy-to-maintain and secure. Sadly, few are: time-to-market of analytics applications is measured in multiple months and, in many cases, they never make it into production. Gartner estimates failure rates for analytic initiatives to be greater than 80%.[8]

Data Debt in The Pipeline Jungle

As the demand for predictive and prescriptive analytics rises this 'data wrangling overhead' is simply unstainable.

Even if an analytics project is successful and is deployed into the enterprise, the pipeline approach is creating problems for the future. Pipelines stored as code may quickly become indecipherable except to the original author. Google describe these as "pipeline jungles"[9] and notes that data dependencies are one of the key contributors to technical debt in machine learning systems and that data dependencies have a higher cost than code dependencies. With high turnover, the average tenure for data scientists is less than a year[10], it is imperative that new recruits can use, adapt and understand the models and model training created by their predecessors.

One of the barriers to achieving this is the plethora of languages favoured by data scientists for the creation of predictive models. There is a perception that data scientists are curious and inquisitive, constantly researching new tools and processes, and expanding their knowledge to encompass the very latest skills and techniques. The current enthusiasm for machine learning and AI has certainly led to a proliferation of new analytic tools, languages and frameworks with which data scientists rush to experiment.

7 https://www.information-age.com/productivity-in-data-science-123482699/ and Dasu, T, & Johnson, T (2003) Exploratory, Data Mining and Data Cleaning (Vol. 479), John Wiley & Sons.

8 https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

9 https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

10 https://www.indeed.co.uk/salaries/data-scientist-Salaries#:~:text=The%20typical%20tenure%20for%20a%20Data%20Scientist%20is%20less%20than%201%20year.

In practice, the data science community is very diverse and different groups tend to stick to the skills they have and focus 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 and the R user will work almost exclusively with the library of scripts they have developed over time. Teams may well be more efficient, comfortable and creative using tools with which they are familiar and attempting to introduce different coding languages and methods could hurt productivity and may be met with opposition.

As the market matures, it is likely that some currently 'hot' tools will fall from favour. Picking 'winners' in this context is fraught with risk and uncertainty.

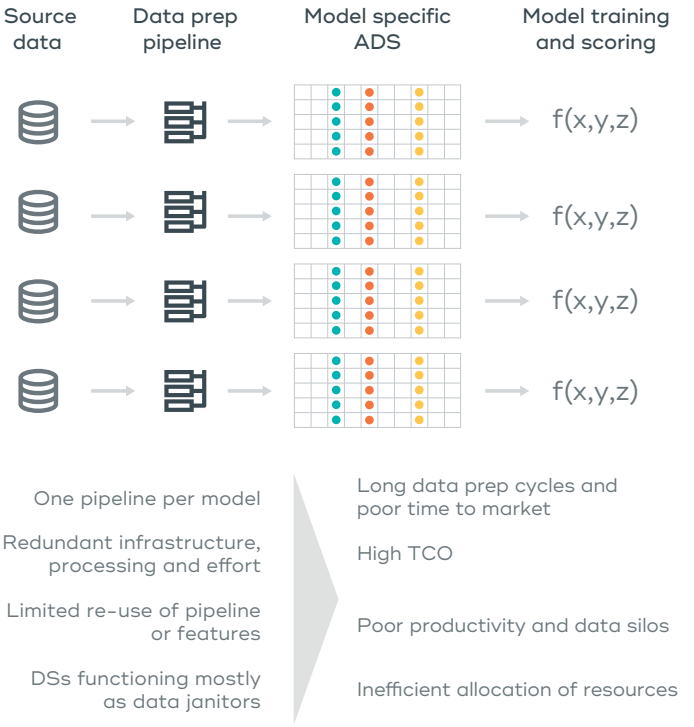
Restricting model training activities to a single technology is also seldom desirable. No single model training tool, language or framework has been able to establish a dominant position in the marketplace. As the market matures, it is likely that some currently 'hot' tools will fall from favour, that others will become niche or be consolidated into broader offerings. Picking 'winners' in this context is fraught with risk and uncertainty. Moreover, there is no single objectively 'best' technology for the wide variety of analytics requirements in large and diverse organisations; a good outcome can be often achieved using multiple libraries, methods, and languages.

Business and analytics team leadership face a dilemma. The pipeline approach to analytics increasingly threatens an organisation's ability to compete today and to transform to meet the demands of tomorrow's digital economy. Existing ways of working, over-emphasis on building and training models, plus poorly optimised approaches to data preparation and industrializing deployment of analytics are all conspiring to slow the effective implementation of AI and machine-learning across business sectors.

At the same time, intensifying scrutiny across industries demands that all machine-learning and AI processes must be auditable and transparent. It is increasingly important, especially in the regulated industries, that organisations are able to understand and demonstrate why and how an AI or machine learning system made a particular prediction from any arbitrary point in the future. Features and predictive models wrapped up and stored as code written by a long-departed data scientist will be impossible to unravel at a later date.

Yet, businesses need to be agile, responsive to changes in their data and the world, as well as flexible to use the latest tools and the skillsets of the best data scientists available. A dynamic 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. Faced with unsustainable project-by-project pipeline development and an inability to standardise on a single analytic tool, language or framework, how can organisations scale the deployment of robust, enterprise-grade analytics?

The one pipeline per model approach



teradata.

Conscious De-Coupling for Focus

The answer lies in 'de-coupling' the various parts of the process and instead focusing on three key components: Feature Engineering; Model Training; and Deployment. A brief description of each element follows, but in totality Analytics 123 seeks to combine freedom with governance by combining the optimal technology for feature engineering, reuse and deployment with a wide choice of tools for model training.

In the same way that modern IT architectures separate storage from compute, decoupling the separate parts of the analytics process leads to a more efficient system and supports the principle of "Polyglot programming"[11] so that appropriate tools, languages and frameworks are applied to tasks to which they are best suited.

It is interesting to see that the organisations which are currently most successful in the use of analytics 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. Re-imagining the feature engineering process to focus on creating reusable features that can both train and score multiple models, significantly reduces the collective time spent on data wrangling. Allowing data scientists to select preferred or best-suited tools to create and train models, confident that they can be imported seamlessly, improves deployment and scoring of live data at scale. The strategy eliminates duplication, removes the need to move data between different platforms and provides robust auditability, monitoring and updating of models at scale.

When considering, implementing and deploying machine-learning and AI within organisations the recent tendency has been to over-focus on creating and training the predictive models that sit at the core of analytics projects. However, although they are seen as the exciting and 'sexy' part of the data scientist's role, in reality they represent only small part of the overall project. The basic tenets of Analytics 123 are that organisations can only successfully scale their machine learning and AI initiatives if they pay greater attention to crucial elements either side of model training—feature reuse and model deployment. Consequently, significant focus should be given to populating and maintaining an Enterprise Feature Store as the foundation of machine-learning and AI across the organisation.

Organisations can only successfully scale their machine learning and AI initiatives if they pay greater attention to crucial elements either side of model training.

The Enterprise Feature Store

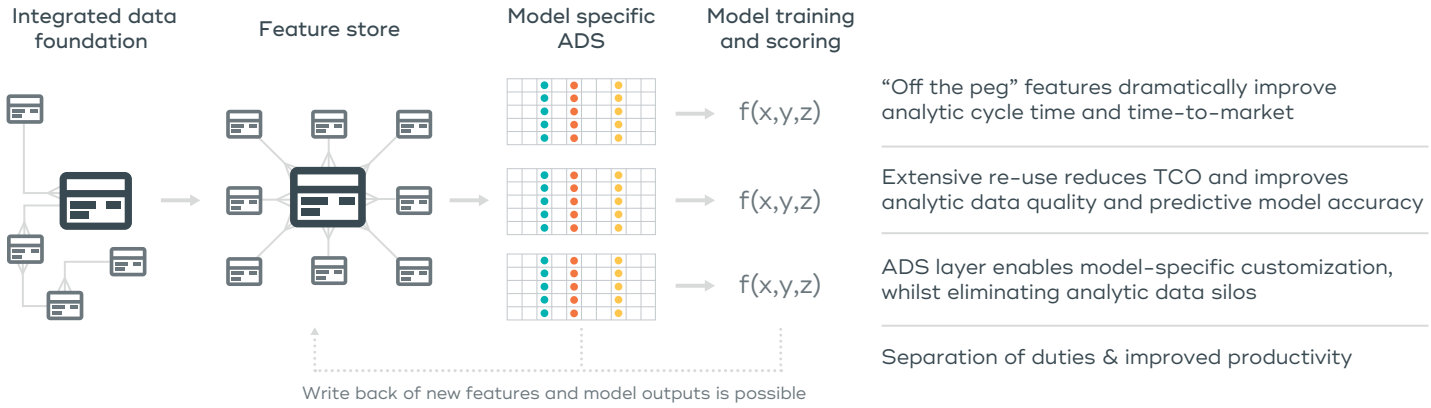
An Enterprise Feature Store (EFS) is a curated collection of variables with proven predictive value, materialized as tables in an analytic RDBMS. The hard, time-consuming work of data preparation, data integration and feature engineering can be done once, creating features that can be reused to both train and score multiple different models. Time and care will need to be taken in creating features which have utility as well as predictive value, and in cataloguing each one, but this initial investment quickly pays of as subsequent projects can easily reuse existing, well-documented features. This is critical—for machine-learning and AI to deliver their promised value they must become ubiquitous, and that means significantly reducing the 80% of cost and effort currently consumed by data preparation and management.

Precise cataloguing of these features, and their ongoing maintenance within the EFS, also contributes to consistency. Different models using common features can be scored with confidence in the validity of predictions. Regular testing and updates can manage model drift.

Enterprise Feature Stores are already dramatically improving the productivity of data scientists and time-to-value for new analytics in leading organisations. Capitalising on Teradata's industry-leading performance and scalability for the manipulation and processing of large analytic data-sets an EFS will also deliver business and cost-advantages by avoiding data movement and duplication, reducing both total cost of ownership (TCO) and latency.

11 http://nealford.com/memeagora/2006/12/05/Polyglot_Programming.html

The feature store approach



The second step, model training, is the home territory of the data scientist, which is perhaps why it has attracted so much focus. Indeed, the dramatic over-emphasis on model creation has contributed to the low productivity of data scientists. It is clear that production analytics at scale does not begin or end with a predictive model and that neglect of essential processes either side of this step are delaying wider deployment.

Wide Choice For Model Training

When it comes to model creation it is likely that for the foreseeable future businesses will need to use multiple different technologies. Smart, agile businesses will wish to think twice before excluding or mandating any specific tool or technology that has the potential to create value. Data scientists must have freedom 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 variety of tools. As noted above, data scientists have their own preferences, different tools are best suited to specific situations, and new tools, languages and approaches are constantly being innovated. Analytics 123 explicitly allows for the use of multiple tools and languages from multiple vendors so that data scientists can select the most appropriate approach for a particular use-case. However, the data required to train the models should come from reuse of variables stored in the feature store and any new features created should be added to the EFS for reuse by others.

Treating model creation as a separate activity allows seamless incorporation of models trained in external systems alongside those created in the database itself.

There is discussion among data scientists as to whether a "model" is the result of an algorithm being trained on data or if it consists of the trained algorithm plus the features that created the training data. With Analytics 123 feature engineering and model training are treated as two separate activities. The iterative nature of model creation means that in the discovery and evaluation phases these two activities are intrinsically linked, but once a model has been created and shown to be accurate, the feature engineering code should be migrated to the feature store and not left tied to a specific model. Treating model creation as a separate activity allows seamless incorporation of models trained in external systems alongside those created in the database itself. Data scientists get the best of both worlds; documented features with proven predictive value; their choice of modelling languages and training platforms; plus, easy portability into Teradata's platform for industry-leading performance and scalability. This 'Bring Your Own Model (BYOM)' approach initiates the critical third phase of Analytics 123 . Ultimately, the value of any analytics project can only be realised when predications are made on live data in ways that can provide timely, actionable business insight.

teradata.

The model training task is typically completed using carefully chosen samples of historic data. By contrast, the model scoring process requires access to complete and up-to-date datasets. It is typically mission critical, requires predictions to be made available at an operational endpoint and is increasingly being executed in near real-time. The challenges of moving from model training to model scoring are often under-estimated and consequently a major cause of failure in analytics projects. The high availability and industry-leading mixed-workload management capabilities of Teradata Vantage avoid many of the challenges organisations face in operationalising analytics.

With BYOM, data scientists can use the most appropriate tool to train any predictive model confident they will be able to score them at scale directly against production data from the Enterprise Feature Store. Tight integration and a variety of methods including PMML, SQL conversion and native code running in-database allow externally trained models to be scored in production in-database and deployed at scale.

Production Delivers Value

The scoring of live production data creates predictions that are used by the business on a regular basis and which creates real value and ROI from machine learning and AI. It is crucial that the production phase of the process is simple and robust. With the EFS and the trained model in place everything required exists in-database; no data movement to or from external systems is required. In addition, Teradata systems are 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 between tens and hundreds of milliseconds. In addition, Teradata's 'always-parallel-all-the time' architecture means that batch scoring workloads are both performant and scalable, so that new predictions can be created on live production data as often as required.

With the EFS and the trained model in place everything required exists in-database; no data movement to or from external systems is required.

Model scoring is typically an example of an "embarrassingly parallel" process and the nature of Teradata's logical, hash-based filesystem means that near real-time scoring operations in Teradata are generally "single AMP, single (logical) IO" operations that 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.

In the near future, organisations will need to deploy hundreds of millions of predictive models in production and support ubiquitous machine learning to remain competitive. Doing so will require a strategic approach to machine leaning and AI that drastically reduces current resource and cost requirements, and speeds time to deployment. Data silos, fragmented systems and duplicated efforts must all be avoided, not only to ensure maximum ROI but to avoid the 'pipeline jungle' that creates 'data debt' and audit nightmares. All of this must rest on the foundations of a scalable and performant data platform.

teradata.

Teradata's always parallel, all-the-time architecture and processing model is perfectly suited to the high-performance processing of large and complex datasets which characterize many data preparation, model training and model scoring tasks. Teradata has proven 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.

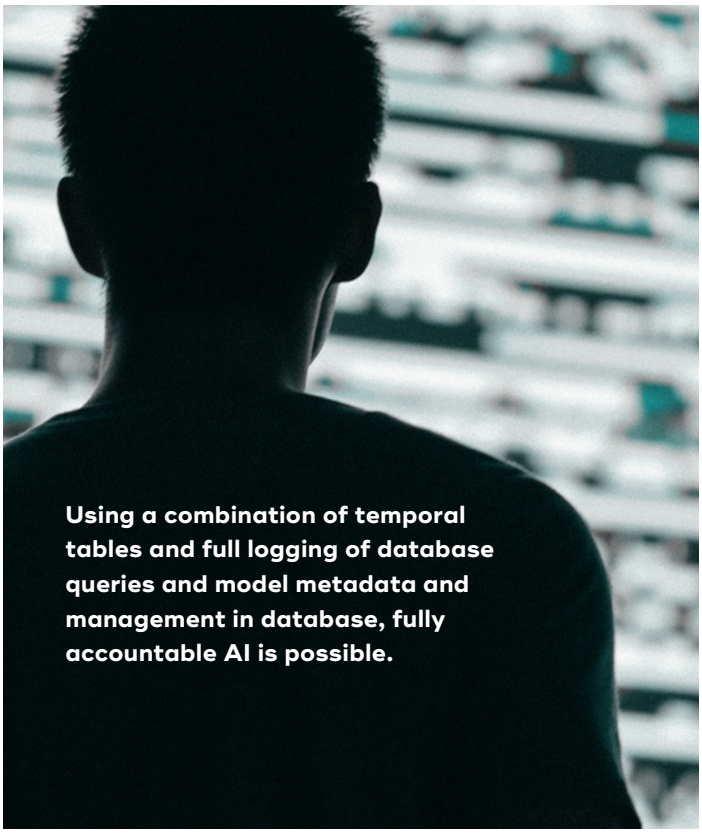
In the near future, organisations will need to deploy hundreds of millions of predictive models in production and support ubiquitous machine learning to remain competitive.

Teradata provides O(1) access to localized data, enabling extremely high throughput and low-latency for tactical queries, like near real-time model scoring. Industry-leading, mixed-workload management capabilities enable these mission-critical operational workloads 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. 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 Teradata database, enabling rapid and flexible data exploration.

All elements of Analytics 123 can be performed in-database. Plus, processing of the analytical languages and libraries popular with data scientists in database eliminates the need for reimplementation of models developed using external tools. Teradata's AnalyticOps accelerator allows full automation of the analytics process including the use of CI/CD pipelines to maintain Enterprise Feature Stores and model management at scale. This means the time to deliver Analytics is greatly reduced.

Teradata provides true hybrid cloud deployment capabilities, including on-premise, virtual private cloud and public cloud platform deployment options. As Teradata is exactly the same product offering irrespective of the underlying infrastructure platform, it eliminates the need to re-engineer existing applications and this ease of portability both lowers the barriers for organisations to adopt a new platform model as well as providing a future exit strategy.

As noted, regulatory scrutiny, and business risk management are demanding increased transparency, documentation, consistency and auditability of machine learning and AI. Using a combination of temporal tables, which allow data to be viewed in the exact state it was at a particular moment in time, and full logging of database queries and model metadata and management in database, fully accountable AI is possible with Teradata. With this information precise investigations of why a particular model made a particular prediction on a certain date can be made, providing for auditability and repeatability.



teradata.

Eighty Per Cent Failure Rates Are no Longer Acceptable

Pipelines are an established way of working for data scientists and have undoubtedly played an important role in the creation of machine-learning and AI capabilities within many organisations. However, they are no longer fit for purpose, and certainly do not support the demands of data-driven businesses of the future. The waste, duplication and sheer resource intensiveness of critical data preparation and integration as part of the feature engineering process must be tackled. Eighty-per cent failure rates and project timelines that extend for months are simply not acceptable and will sink any ambitions to implement AI or machine-learning at a scale where it can generate real value. Approaches that are already struggling today to support the deployment of thousands of predictive models will crumble when faced with the predicted requirement to deploy and score tens and hundreds of millions of predictive models in leading businesses in the next few years.

Analytics 123 presents an alternative. By de-coupling the process into three phases, it ensures that each is given the weight and focus it needs. Features are engineered to be reused, documented and catalogued in an Enterprise Feature Store reducing duplication, increasing efficiency and consistency. Data scientists are freed to use the tools and languages they feel are best for each specific task—knowing that trained models can be easily ingested back into the enterprise to score live data in the Enterprise Feature Store. And that scoring can leverage the massively parallel, high-performance, enterprise scale capabilities of Teradata to drive real-world, business-critical analytics that can transform organisations.

About Teradata

With all the investments made in analytics, it's time to stop buying into partial solutions that overpromise and underdeliver. It's time to invest in answers. Only Teradata leverages all of the data, all of the time, so you can analyze anything, deploy anywhere, and deliver analytics that matter most to your business. And we do it on-premises, in the cloud, or anywhere in between. We call this pervasive data intelligence. It's the answer to the complexity, cost and inadequacy of today's analytics. And how we transform how businesses work and people live through the power of data. Get the answer at **Teradata.com**.

About the Authors

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.

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

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

teradata.