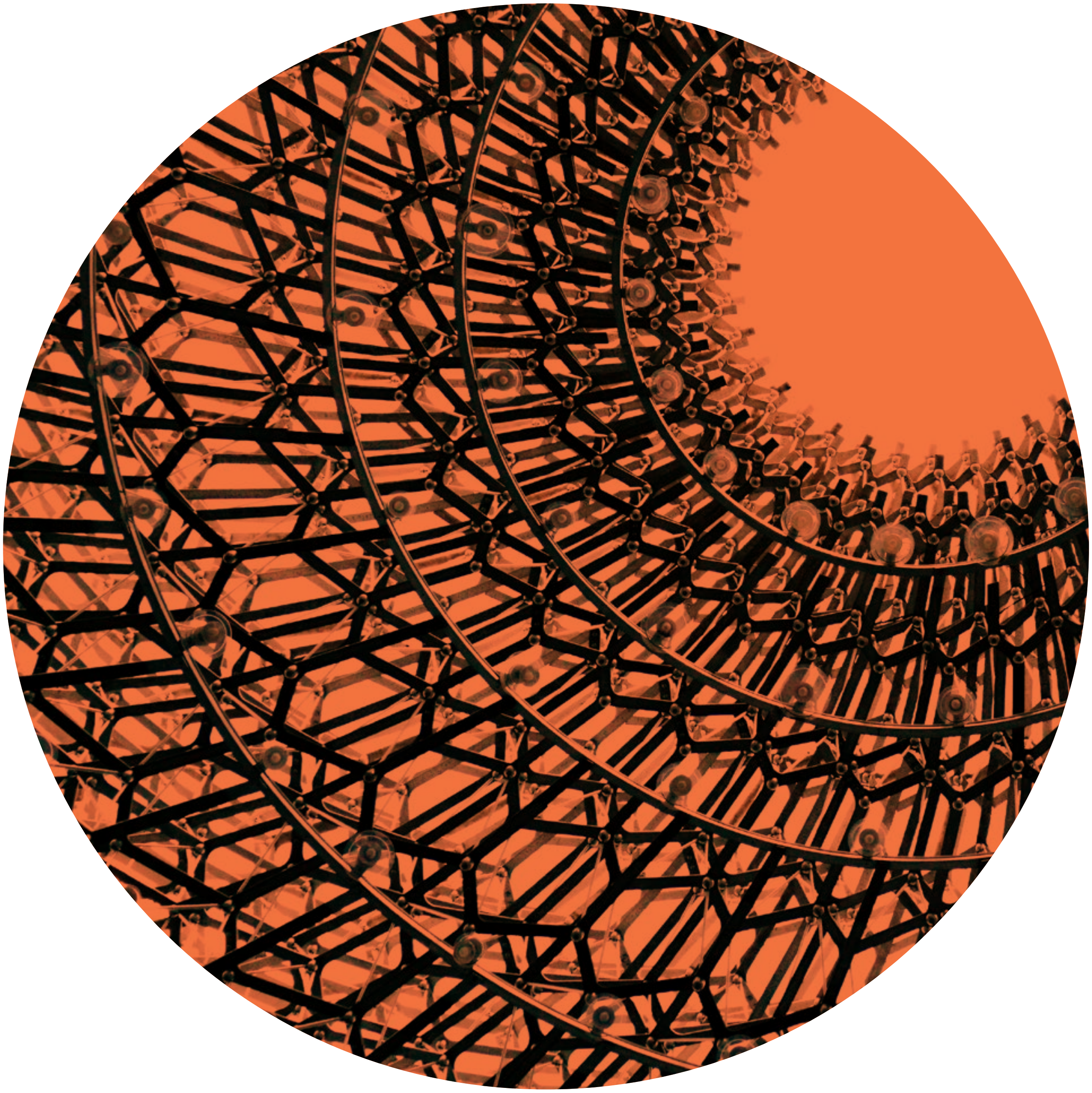
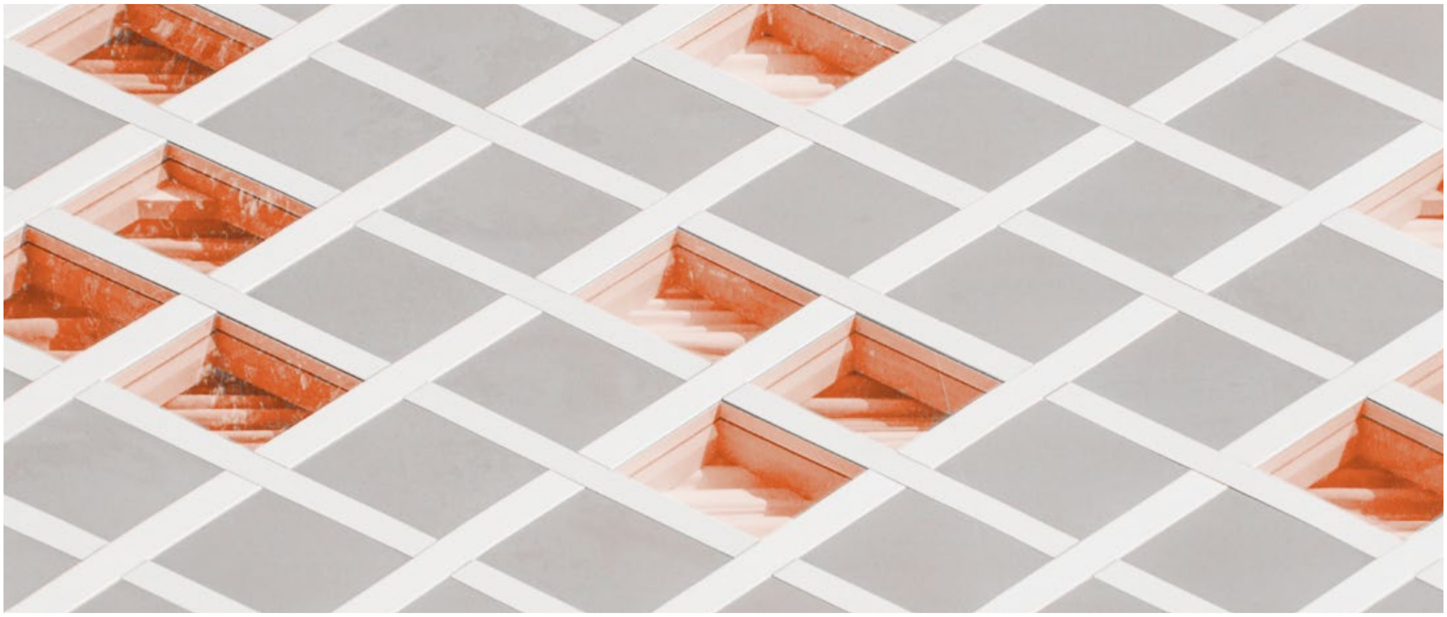
WHITE PAPER

Throughput Review: Distributing the Process



Matt Reubendale, Business Analytics Leader, Teradata
Karen Diamond, Business Analytics Engineer, Teradata
Cheryl Wiebe, Industrial Intelligence Practice Director, Teradata

12.18 / EB10273 / WHITE PAPER



How the Teradata Analytic Framework uses a unique Teradata capability:

Key Points

Teradata® Vantage, the platform for Pervasive Data Intelligence

Teradata Vantage delivers the speed, scale, and flexibility to deliver genuine answers. Algorithms, AI, ML, and every other analytic process is just math—and Teradata is extremely good at math. If you know how to take advantage of its unique capabilities, like reframing the problem so that it can be broken down and done in parallel, and controlling that parallelism, Vantage can help you solve problems that seem too large to tackle.

Exporting is Expensive

Exporting data from Teradata is the least efficient way to use the resource. It's possible to do what you're doing in the database and it will likely consume significantly fewer resources and take less time than exporting.

On Teradata, SQL is the Most Efficient Language

It is possible to use Python, R, Spark, and other languages on Teradata; however, it is often inefficient, especially on the largest systems. Understanding when a process would benefit from being converted to SQL and run in-database can yield significant analytic throughput improvement.

Teradata is Uniquely Capable of Distributing a Process

Parallelism is not unique. In fact, Teradata competitors will highlight that you don't have to manage their parallelism. But the capability to manage that parallelism enables Teradata users to distribute processes, not just data.

Teradata Consulting Can Help

Teradata delivers a proven platform, backed by a rich 40-year history of evolving analytic capabilities. Teradata Vantage combines descriptive, predictive, and prescriptive analytics that fuel AI and Machine Learning to activate data intelligence across the enterprise. Supported by Teradata Consulting—more than 5,000 global experts in Pervasive Data Intelligence—we leverage transformative technologies and our unique expertise to help you build massively efficient analytics, and learn how to take advantage of these powerfully unique capabilities to answer questions previously deemed unanswerable.

Summary of Concept

Shared-Nothing Massively Parallel Architecture

Early on, Teradata made the decision to use a shared-nothing, massively parallel processing architecture. This design decision is what enables Teradata scalability, as each AMP (virtual compute unit) owns an equal slice of the disk. Only that one AMP reads that one slice.

By distributing the data across the AMPs, this architecture enables any query of that data to be extremely fast and efficient. An "embarrassingly parallel" query like a sum or a count is ideal for this architecture. For example, if we had 1,000 random playing cards and we wanted to know how many kings there were, the fastest and most efficient way to answer that question would be to split up all the cards across all the units of parallelism and ask, "How many kings do you have?" Each unit of parallelism would report back and then they would be added together.

Obviously, sums, counts and other embarrassingly parallel queries do not answer all of the business questions we need to answer. In fact, the great majority of our business questions today require combinations of various algorithms, distributions, and other math that is changing every day. Many of these analytic capabilities are developed as open source projects or are available through applications (e.g., R or Weibull ++) or libraries (e.g., Cran, PyTorch or Turi).

The perception is that the only path to be able to leverage these emerging capabilities is to export the data from Teradata to other analytic platforms, but that's not the case. In fact, it is possible to leverage these capabilities so that the data never has to move, which significantly reduces the time it takes to complete an analytic (i.e., increase your overall analytic throughput) and reduces the resource consumption on Teradata.

Exporting is the Most Expensive Step

In regards to time, resources, and complexity, the most expensive and least value adding step in an analytic process is exporting from the enterprise data warehouse. Teradata wasn't designed to simply be a storage layer; rather, a powerful analytics platform designed to perform the most complex analytics at scale. If it's not being utilized in that way you're likely introducing complexity and cost unnecessarily into your analytic process. This has the direct impact of reducing productivity—for you and your organization.

SQL is Dead—Long Live SQL

Did you know that you can run Python, R, and other emerging languages on Teradata? Wouldn't it be great to simply copy and paste that new algorithm—the one that perfectly explains your training data set—and just have it run at scale against all of the data in milliseconds? We think that would be great too; however, it's pretty inefficient—and especially risky in the largest Teradata systems.

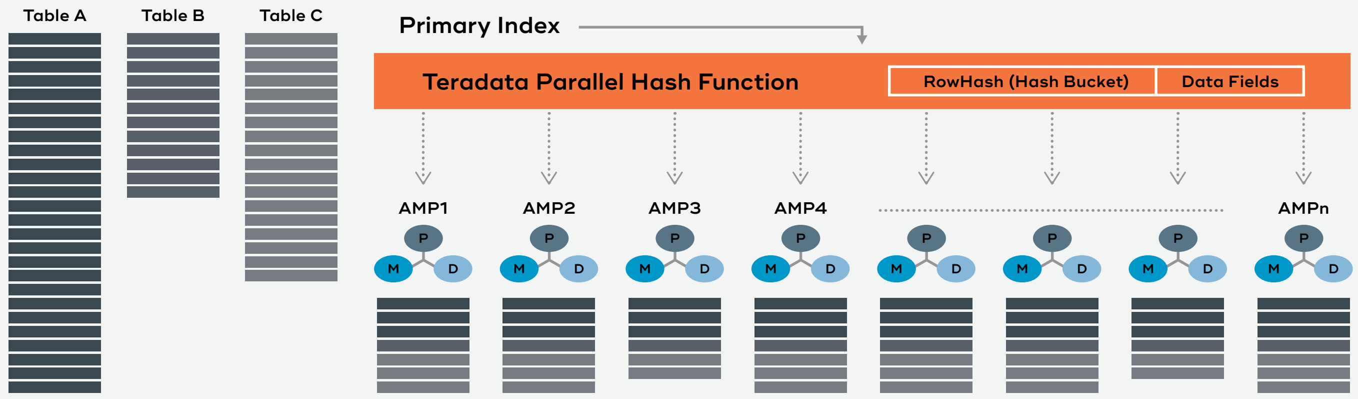


Figure 1. Teradata shared-nothing architecture.

The Teradata Optimizer was developed to run SQL since 1979. You likely don't know the Optimizer—you've never seen it or asked for its help—but if you've used Teradata you've benefited from it. It reviews every single query against Teradata and turns the query into a series of steps for execution. Those steps might include moving data between AMPs, avoiding a whole table scan, changing the Primary Index of a table, or even rewriting your query (don't worry about that Oracle background, the Optimizer has got you covered). Python, R, and the other emerging languages run below the database layer, bypassing the optimizer and the efficiencies it offers. Teradata continues to invest in making this easier and more efficient with each new version of Teradata Vantage.

Procedural Processing vs. Set Processing

Procedural Processing—The Current Way

If you can't simply copy a Turi algorithm into Teradata—and even if you did, it likely wouldn't be an "embarrassingly parallel" process—why (other than the export) would you "do more math in Teradata?" Great question, because distributing the data isn't the only thing you can do with the Teradata architecture, which enables us to parallelize processes that don't lend themselves to data parallelization.

For example, let's imagine you're trying to model the effect of a promotional price across five different products, in 50 different US States, and with all 210 Nielson Designated Market Areas (DMAs). These hierarchy levels represent 1,300 different models that need to be created. Each of these models requires a baseline, or what the sales would have been if there wasn't a promotion. Let's assume to create the baseline you want using a simple exponential smoothing algorithm:

$$s_0 = x_0$$

$$s_t = \alpha x_t + (1 - \alpha)s_{t-1},\ t > 0$$

where α is the *smoothing factor*, and $0 < \alpha < 1$.

The two parameters you control are the number of smoothing passes and the a (alpha) which can be any number between zero and 1.

How you might do this today is to export the data from Teradata to SAS, R, or another analytic platform and incrementally change the number of smoothing passes and your a until the baseline looks right or has a high correlation with your training data. Now you just have to do this 1,299 more times and you have a thorough understanding of how the promotional price affected sales.

Distributing the Process

Teradata users, uniquely, can control how the Optimizer parallelizes a query which means that Teradata, uniquely, can distribute a process—not just the data. How this capability impacts the example above is that, instead of procedurally trying different a and number of smoothing passes, the Teradata user can calculate all of the a and all of the smoothing passes, for all levels of the hierarchy. The resulting answer set is obviously enormous, a factorial of 1,300 * variations of a * variations of passes, but it can be can be mathematically determined which parameters yielded

Figure 2. Do as much math on Teradata as possible, before handing off to other processes.

the most accurate baseline models for each level of the hierarchy. In practice we've seen users of this capability go from thousands of hours to answer a business questions to 13 seconds.

What's even more impactful about leveraging this capability is that it actually uses Teradata less than an export to another platform. In practice, Teradata users have been able to generate 62,000 more calculations than their previous procedural process while consuming only 2% of the CPU of the export required for that previous process.

It's faster, it's cheaper, and it delivers better answers.

How to Get Started

At this point, hopefully you're rethinking how you ask questions—from, "How did my promotional price drive demand in Denver?" to, "What drove demand?"—and how you can get started.

This capability's value scales directly with the size of your Teradata system. The more units of parallelism you have, the more answers you can generate in a single distributed process. However, it can also increase the impact of inefficiency. Executed poorly, this capability can have a significantly negative impact on resource consumption in Teradata leading to a nasty call from your friendly DBA.

To capture the productivity gains while simultaneously maintaining—or even reducing—your consumption of Teradata resources, we recommend engaging Teradata Consulting. They're available and ready to help implement the Teradata Analytic Framework, which will enable you to take advantage of its unique capability effectively and efficiently. They can also help translate or recreate your preferred algorithms, distributions, and other emerging analytic capabilities into Teradata SQL that will run accurately, efficiently, at scale, and within the Teradata Analytic Framework.

Definition of Analytic Throughput

Analytic throughput is a quantifiable measure of the count of business outcome-related answers an analytic process can deliver in a specific amount of time. By defining the measure in this way, analytics organizations are able to quantify the value of improving the throughput of a singular process, or of their entire organization. Our assumption is that for every quartering of the time interval required to provide an answer, the productivity of labor and of working capital can often double. These productivity gains can result in as much as a 20 percent reduction in costs overall (Stalk Jr & Hout, 1990).

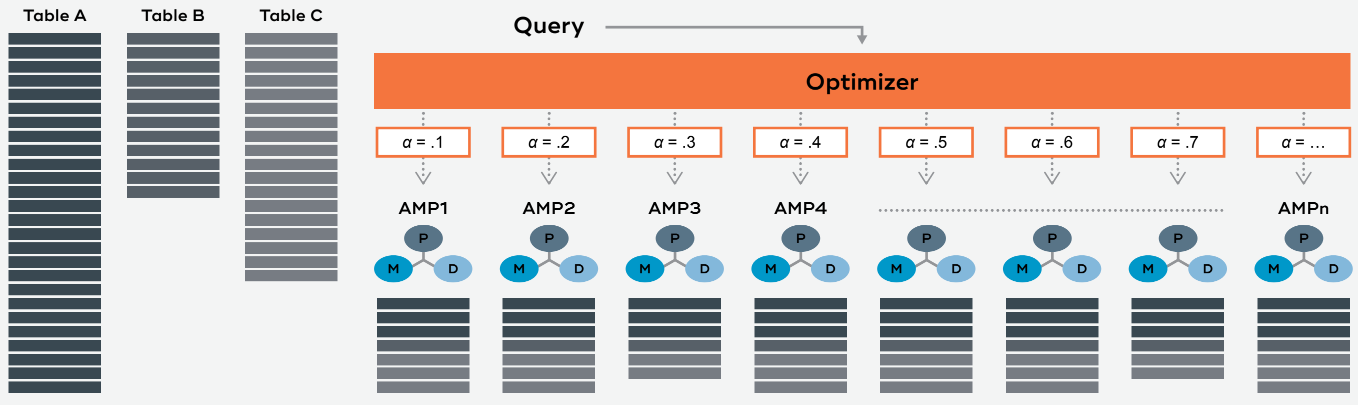


Figure 3. Using the Optimizer to distribute a process—and the data.

Teradata Business Analytics Team

Our knowledgeable business analytics team is dedicated to helping business users capture as much value for their organization as possible from their investment in Teradata technology. To this end we work directly with business users, in partnership with their IT organization, to review their analytic processes and provide recommendations for how to improve their analytic throughput.

About the Authors

Matt Reubendale, Business Analytics Leader

By focusing on how analytics affects decision making, Matt has had the opportunity to work with myriad organizations on how to leverage traditional and emerging analytic platforms to transform their businesses. He and his family relocated from Minneapolis to San Jose in 2017.

Karen Diamond, Business Analytics Engineer

Applying hard-won experience from her background at Delta, Coca Cola, and dozens of Teradata customers, Karen combines deep technical knowledge with a first principles approach. Karen relocated from Atlanta to San Francisco in 2016.

Cheryl Wiebe, Industrial Intelligence Practice Director

As data volumes continue to grow, Cheryl has found her organizations expertise in analyzing immense data sets with speed and precision more critical than ever to solve the most pressing data challenges. Cheryl and her husband live in Southern California and visit their daughters often in Minneapolis.

About Teradata

Teradata leverages all of the data, all of the time, so you can analyze anything, deploy anywhere, and deliver analytics that matter. By providing answers to the complexity, cost and inadequacy of today's analytics, Teradata is transforming how businesses work and people live.

For more information:

Contact Matt Reubendale at **(612) 208-5282**, email **Matthew.Reubendale@teradata.com**, or visit **Teradata.com**.

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