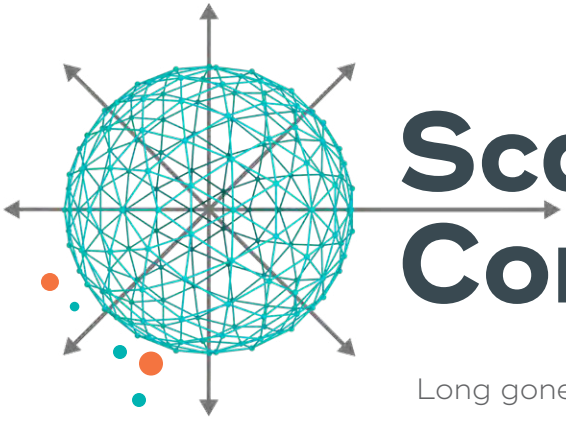


Scale Without Compromise

Long gone are the days when a successful digital strategy meant amassing more data than the competition. Today a winning data-driven approach means knowing what to do with your data—how to derive intelligence that helps you innovate faster than your competitors do. The increasing gap between the digital "haves" and "have-nots" comes down to who knows how to use data to predict the future.

teradata.

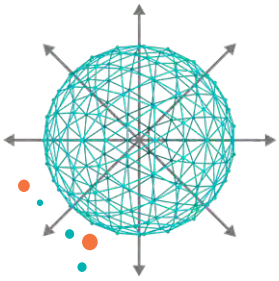
Table of Contents

Needless to say, predicting the future is hard to do. The most innovative companies do so by integrating data across organisational and functional boundaries so that they can optimise complex end-to-end value chains—and using that continuously updated data platform to continually hone. In the era of mass-personalization, customers expect a two-way feedback loop; customers want brands to know them and their exact needs and preferences while also giving them the chance to provide input and take control of their experiences.

The modern enterprise constructs a data analytics strategy that adapts dynamically to changing business requirements and empowers employees, partners, suppliers, and potentially even customers to ask complex data-driven questions. You can't know in advance all the questions that will need to be asked, what data will be needed to answer them, and how many people will be asking them. So you'll need a data analytics platform that supports users and queries numbering in the millions, every single day. As that data analytics platform is increasingly integrated with mission-critical, operational processes so that accurate and up-to-date predictions are available at the customer moment-of-truth, three, four, or even five-nines availability is required—and at Petabyte scale.

Supporting this level of scalability requires the following key capabilities: separation of compute and storage, with elastic scaling; integration with first party cloud services; ingestion of modern data sources; integrated data management and scalable analytics; and dynamic resource allocation and workload management.

As your users become aware of the possibilities to access and leverage data that you're creating for them, they will ask increasingly sophisticated and complicated business questions. These more complex business questions will likely result in high levels of deeper insight and return on investment as your people bring to light information that was previously unavailable. Creating a system that scales with these fluctuating and ever-evolving demands will be key to gaining a competitive advantage.



teradata.

Uncertainty Can Be Expensive

Predicting the future—whether that's the customer's latest preferences, the possibility of a machine part failing, or the next opportunity to cut costs in supply—has become more complex in a constantly changing environment that's innately uncertain. And uncertainty can be expensive. Not being able to predict the impact that adding hardware or users will cause leads to costly over-configurations with plenty of unused capacity. Or uncertainty can lead to under-configured systems that can't meet required performance levels, resulting in expensive redesigns.

Such uncertainty makes scaling a data analytics platform potentially highly costly as well. Many vendors and industry professionals define a scalable system as one that allows multiple queries to be run simultaneously and that can expand in the case of future data growth needs. But such a vague definition of scalability simply means that the platform is capable of some growth and expansion, with no discussion of the granularity of this expansion or the impact that this expansion has on factors like latency, performance, reliability, availability, and, critically, TCO.

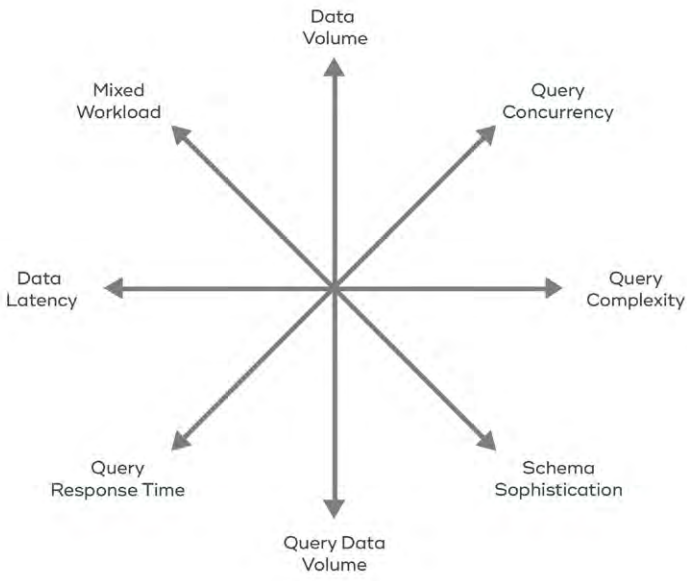


Figure 1: Multidimensional scalability of data analytics platforms requires increasing capabilities across eight dimensions simultaneously, without adverse impacts on either domain.

The Eight Dimensions of Scalability

A more accurate definition of scalability considers the eight domains that are vital for any data analytics system. These are shown in Figure 1 and listed below:

1. **Data Volume** refers to the ability to efficiently store and process petabytes of raw user data in the same system, natively and in object storage. This makes it easier for the business to have access to all the data they need to drive deep insight. By leveraging Cloud object stores like AWS S3, Azure BLOB and GCP Object Store, Native integration, users can read directly from these stores in a high-performance way.
2. **Query Concurrency** refers to the ability of the system to simultaneously process large volumes of queries and to get more work done, faster. This enables optimization and balance of numerous complex, resource-intensive queries while maintaining SLAs. As demand requires more query concurrency, you can scale out with the cloud dynamically.
3. **Query Complexity** is the ability of a system to support complex, high-value queries, including multi-join queries. Users often ask questions that span different functions in a business. These questions often generate very sophisticated queries, so the ability to handle such queries empowers the business to uncover insights that were not previously known. As the complexity of the queries increases, you can scale up with the cloud by adding more powerful compute to handle without redistribution.
4. **Schema Sophistication** refers to extensible and flexible data schemas that support and enable all business requirements. Businesses rely on a variety of schemas, such as normalized schemas, to discover complex relationships between data entities and unlock insight. But the enterprise relies on the ability to support any schema—normalized, semi-structured, or no schema at all—to meet the needs of the business. This dimension delivers the flexibility and agility to realize rapid time-to-value. Support for native integration with object store data types (AVRO, Parquet, JSON) enables seamless joining with production data.

teradata.

5. **Query Data Volume** is the volume of data that can be processed quickly and efficiently by a single query, without manual intervention. This capability prevents business users from having to limit their queries—so they can consider all of the enterprise's data in their queries if needed—and optimizes I/O throughput for enhanced performance. When you scale up with the cloud, you can deliver more memory and compute to handle larger query volumes.
6. **Query Response Time** is the ability to deliver fast and consistent response times to comply with strict SLAs. This dimension contributes to agile innovation at the enterprise. Query response time determines not only the performance of queries that are known questions but also those that are unknown, unanticipated, exploratory, broad, or complex as well as those that are associated with new applications. Native integration with object stores enables users to seamlessly query data across the persisted data in high performance SSDs and the lower performance object storage.
7. **Data Latency** is the ability to load and update data in near real-time while simultaneously supporting query workloads. This is an indication of whether the data in the warehouse can be kept stable, current, and in sync with business processes and operations to the degree necessary to respond to events and business needs as well as to provide meaningful analyses. The cloud enables you to stream via an Event Driven Architecture with first party cloud services from Azure Event Hub, AWS Kinesis and GCP Cloud Data Flow.
8. **Mixed Workload** refers to the ability to support multiple applications and users with very different SLAs in a single environment. Being able to manage multiple workloads concurrently on the same physical platform simplifies management, helps the business guarantee SLAs, and maximizes resource efficiencies. Scaling out and up with the cloud will deliver extra compute that can be directed to high priority workloads, providing dynamic resource allocation.

Just as you can't predict all the questions that will need to be asked and the data that will be needed to find the answers, you cannot predict which data analytics dimensions you'll need to scale when. So you need to be able to scale simultaneously, across every dimension. Focusing on a limited number of dimensions at a time could diminish the capability of any of the others or cause you to miss an opportunity to leverage data and garner additional insights.

For example, as projects get started, they are typically designed for an initial static environment that is represented in Figure 2. If data size is important, the project may even be able to tune for a large data volume or size.

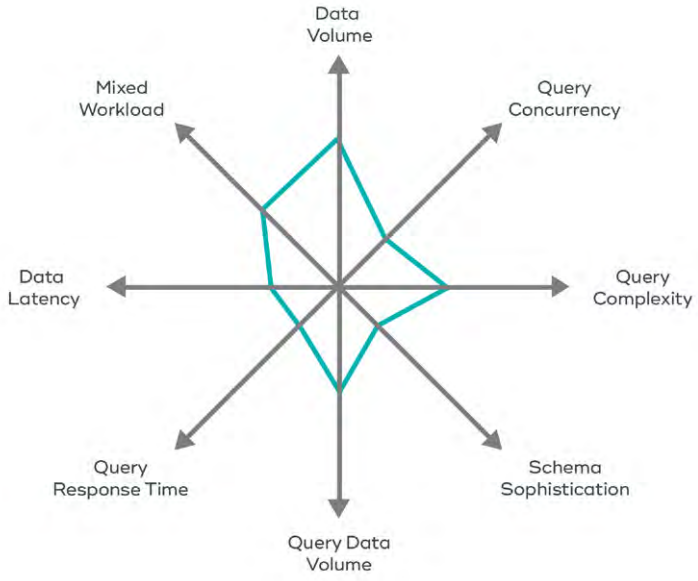


Figure 2: As projects get started, they are typically designed for an initial static environment.

Successful projects will drive demand for broader use by a wider number of users, which in turn will increase the need for query concurrency. And as users become aware of the potential they will ask increasingly sophisticated and complex business questions, which will require greater query complexity.

But technologies are often unable to handle increasing numbers of concurrent, complex queries against larger volumes of data. Thus, to meet requirements of a dimension such as query complexity, the enterprise often must compromise in other dimensions (See Figure 3). In order to handle a significant increase in the

teradata.

number of concurrent queries to meet their business needs, they may need to reduce the complexity of their complex queries to maintain their performance levels. Or they will need to implement separate platforms—one that will handle larger numbers of queries that are highly constrained and tuned, and a separate system tuned for complex queries but with limited users. This adds cost and complexity to the environment.

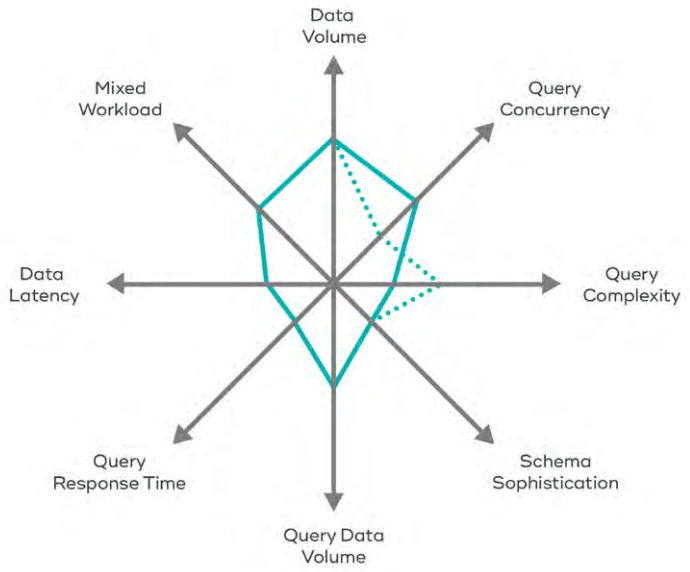


Figure 3: Technologies are often unable to handle increasing numbers of concurrent, complex queries against larger volumes of data. Thus, they must compromise in other dimensions.

Multidimensional Scalability is Achievable

The modern enterprise must embrace multidimensional scalability, where all data analytics domains are enhanced together. Where expansion of processing power leads to proportional performance increases. Where overall throughput is sustained when adding users, and data growth has a predictable impact on query times. This approach future-proofs the enterprise, ensuring that investment in one dimension strengthens the capabilities in other areas and supports the massive workload needs of tomorrow, today.

In the previous example, query concurrency and complexity could not be scaled simultaneously. Teradata Vantage addresses this challenge by optimizing both of these capabilities so that they can scale together.

First, Vantage's "all parallel, all the time" architecture ensures that the four most expensive data platform operations—scans, joins, aggregations, and sorts—always run with the maximum degree of parallelism possible. This dramatically improves performance and throughput for these operations. Vantage also combines the industry's leading cost-based optimizer, hash-based indexes, aggregate join indexes, and support for a wide-range of join plans. Between them, these capabilities enable the efficient, high-performance execution of complex queries, without the need for manual intervention like database "hints" and query re-writes.

Finally, Vantage's Incremental Planning and Enhancement (IPE) optimizer extensions mean that complex queries can be optimized even where good database statistics are not available and where data are not stored locally on the filesystem.

When it comes to supporting query concurrency, Vantage's virtualized parallel RDBMS and BYNET networking protocols ensure that all operations and steps run with the maximum degree of parallelism possible, so that operations execute quickly and efficiently and that work is processed as quickly as possible, maximizing both concurrency and throughput. A logical hash-based filesystem supports efficient, high-performance O(1) localized data access with, so that "tactical" queries are essentially free in Vantage and do not consume significant system resources or block other queries. In addition, Vantage's sophisticated mixed-workload management allows the system to manage extreme levels of concurrency. In fact, several Vantage customers regularly run several thousand concurrent queries on a single Teradata system.

Thus, as shown in Figure 4, Teradata empowers the customer to scale query concurrency and complexity at the same time, without compromising the capabilities of either or dragging down performance.

teradata.

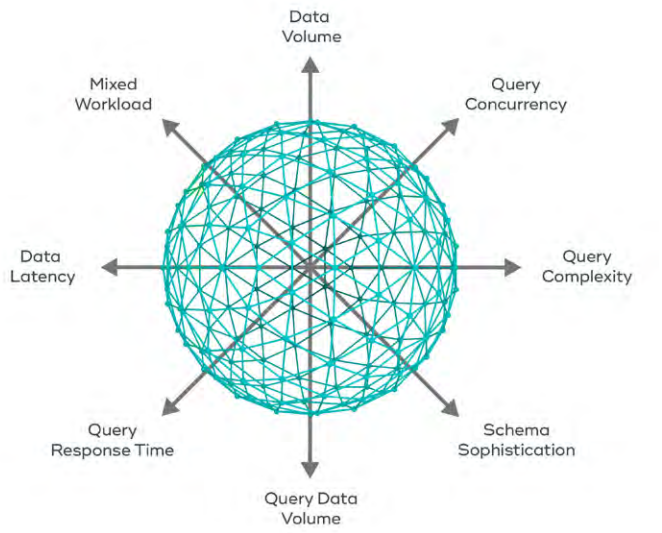


Figure 4: With Teradata Vantage, the enterprise can scale across all eight key dimensions of analytics at once, preventing the need to compromise on performance in any area.

Refuse to Accept System Technology Limitations

Technology choice should not be a limiting factor in the enterprise's ability to scale—but data analytics technology does need to meet specific criteria. As Figure 5 shows, there are clear specifications for achieving multidimensional scalability. Technology needs to support a 5 on all 8 of the dimensions listed.

According to these requirements, multidimensional scalability is achieved by:

1. Efficiently storing and processing over 10 petabytes of data
2. Supporting complex extensible and flexible schemas, including 3NF
3. Handling queries with up to 64 joins
4. Managing dynamic mixed workloads
5. Allowing a single query to access all stored data if required
6. Running over 1,000 queries concurrently
7. Loading and updating data in operational real time
8. Delivering query response times that are consistent against varying workloads

We created this matrix to show that multi-dimensional scalability is achievable—and we've proven so with Teradata Vantage. Our platform delivers capabilities in the "5" column listed in Figure 5, across all dimensions:

1. **Data Volume**—The largest Teradata system currently in production manages in excess of 10 PB of data, and Teradata frequently supports tables that have trillions of rows. By combining a coordinator node-free MPP architecture with a sophisticated cost-based optimizer and hash-based index structures, Vantage delivers "always-parallel, all of the time" data processing.
2. **Query Concurrency**—With virtualized parallel RDBMS and BYNET networking protocol and sophisticated mixed-workload management, Vantage ensures the highest level of parallelism possible. Users are able to bring more applications, users and queries to a single platform, with cost-per-queries that are commonly measured in hundredths of cents. Vantage supports systems today that are running 100+ Million queries per day.
3. **Query Complexity**—With an industry-leading cost-based optimizer, hash-based indexes, aggregate join indexes, and support for a wide-range of join plans, Vantage users can execute complex queries efficiently and at high performance, without having to make manual adjustments.
4. **Schema Sophistication**—Vantage supports complex, nested schemas and views, enabling simplified representations of complex data to be presented to users, tools, and apps without the creation of additional ETL tech debt.
5. **Query Data Volume**—By leveraging advanced indexing and partitioning whenever possible, Vantage's structures are simple to deploy and automatically maintained. Vantage also uses multiple automatic memory caches. Users who can't leverage index structures and caches can use Vantage's advanced compression options, efficient row storage with variable length blocks, cylinder reads, and synchronized-scanning.
6. **Query Response Time**—Vantage's hash-based filesystem ensures that data is evenly distributed which provides the first Index (PI) at no extra cost and delivers tactical queries at an O(1) complexity.

teradata.

Cost Based Optimizer ensures that the most optimal query plans are generated regardless of the underlying schema. Vantage also supports indexes that are automatically maintained, which further improves query performance. Vantage's Workload Management provides advanced controls that deliver consistent query response times.

7. **Data Latency**—Vantage's load utilities enable high-performance batch and stream loading of data and features like Load Isolation, Locking controls, MVCC, and more enable the simultaneous querying and loading of data with very low latency. Multiple load jobs can load data to the same table at the same time while remaining separated from one another.
8. **Mixed Workload**—Vantage's Workload Management combines filters and throttles with sophisticated and dynamic allocation of resources. This allows prioritization of urgent, high-priority workloads and protects the system against the impact of "run-away" queries to improve query throughput and consistency so more work can be accomplished.

Figure 5: This matrix scores the extent of scalability for each of the eight key analytics dimensions, with 5 being the strongest ranking.

	Units	1	2	3	4	5
Data Volume	Raw User Data	< 10TB	< 100TB	< 1PB	1–10 PB	> 10PB
Query Concurrency	# of Users	< 5	< 10	< 100	100–1000	> 1000
Query Complexity	# of Joins	< 2	< 5	< 16	16–64	64
Schema Sophistication	Relational. Flexibility	Flat files with non-atomic nested structures	Flat 1NF Files	Limited Dimension Tables	Star/ Snowflake Schemas	Complex 3NF
Query Data Volume	Data Volume	Only cached data	Indexed Data only	Data partitions only	All Stored data with performance penalty	All Stored Data
Query Response Time	Response Time	Inconsistent	Consistent against cached data	Consistent against known queries	Consistent against few ad-hoc queries	Consistent against varying workload
Data Latency	Descriptive	Read/Write isolated	Eventual Consistency	Limited write capacity while reading	Near Real Time	Operational Real Time
Mixed Workload	Descriptive	No WLM	Basic Resource Management	Static Workload Management	Static Workload Management with throttles	Dynamic Mixed Workload

teradata.

Perform While You Transform

With today's race to gain actionable, predictive intelligence from data accelerating, the enterprise must scale up its analytics capabilities at a rapid rate. But for most analytics platforms, scaling is interdependent, leading to difficult tradeoffs in cost and performance: supporting more queries, for example, can mean a significant degredation in data availability; supporting more concurrent queries can result in worse query performance, limiting throughput and compromising SLAs. And many data analytic platforms scale only in coarse increments, creating cost pressures and resulting in high—and unanticipated—charges that train users not to run the complex, exploratory workloads that are required to create new business value.

At Teradata, we accept no such excuses—when it comes to data analytics solutions, we believe that the enterprise should not be limited by technology choice.

The enterprise doesn't have to compromise with Vantage, the only enterprise-grade data analytics platform that lets you scale all dimensions simultaneously. This empowers the enterprise to handle the massive, complex data workloads of the future, today. From queries to users to volume, as demand against one vector increases, performance isn't lost in others. Vantage empowers the enterprise to waste no investment in the future and perform while it transforms.

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**.

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

The Teradata logo is a trademark, and Teradata is a registered trademark of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or Teradata.com for more information.

© 2020 Teradata Corporation All Rights Reserved. Produced in U.S.A. 08.20