CITO Research

Advancing the craft of technology leadership

Cutting through the Complexity of Complex Analytics

SPONSORED BY

TERADATA

CONTENTS

Introduction

Big data changed the analytics game. It's not a question of ***whether*** but ***how*** to integrate big data analytics into daily operations. The range of data management and analytics engines continues to expand rapidly. Outside of the enterprise data warehouse (EDW), new data management and analytics engines offer a plethora of features at various price points. Each engine is a natural home for particular types of data and analytics, enabling us to gain insights and value from data that was previously out of reach.

The difficulty is that each of these analytics engines is a silo. The insights we want require multiple engines. We have analytics from Hadoop, from a graph analytics engine, from an in-memory database, and from NoSQL, each requiring different approaches. Using multiple engines is not easy; it entails moving and transforming data with help from an overworked IT department. Businesses need a simple way to use all their analytics resources.

Businesses need a simple way to use all their analytics resources.

Understanding the capabilities each engine offers is crucial to ensuring that companies know what each engine can do and how to maximize its capabilities. This CITO Research white paper lays out the range of choices and their capabilities and points to new engineering advances that enable companies to harness the capabilities of each engine while cutting through the complexity associated with data movement and heterogeneous interfaces.

So Many Choices, So Little Time

Historically, there has always been a diversity of analytics techniques. We have numerous choices when it comes to analytics these days. You could say it's an embarrassment of riches.

There are choices of:

- Analytics engines
- Memory and disk equipping those engines
- Design patterns

And of course, all of these options are subject to the ultimate constraint: economics. We have to choose the analytics techniques and technologies that we think will offer the best return on investment for our particular use case and business problem.

Engine Choices

The following table offers an overview of some of the available analytics engines.

Type of engine	What it does	Examples
Massively parallel processing relational database management systems (MPP RDBMS)	Engines built on the relational database model where data is organized into rows and/or columns and typically accessed using SQL. The most highly performant databases in this category leverage massively parallel processing and are referred to as MPP RDBMS. Most data warehouses fall into this category.	Teradata®, Oracle Exadata, IBM Netezza, and others
General purpose file systems	Can store all kinds of data, affordably, retaining its original structure. Such repositories can store massive amounts of data on commodity hardware, often in a cluster-style configuration (which helps with both scalability—the ability to add nodes as needed—and fault-tolerance—the ability to replace one node if another one fails).	Hadoop distributions including Hortonworks, Cloudera, and others
Graph databases	Graph databases store data in relation to other data, making it possible to find out how one node is related to others. Consider a graph of call records, in which the caller is a node and the calls placed are edges reaching out to other nodes—the people he or she called.	Teradata Aster®, Giraph, Neo4J, and others
Columnar databases	Columnar databases (and there are a variety of architectures for such databases) are good for computing aggregates over large numbers of similar items. They can return smaller subsets of data when values are extensively repeated.	Teradata, HP Vertica, and others
NoSQL databases	Comprised of document databases, wide column stores, and key-value stores, these non-relational, distributed, horizontally scalable environments have emerged with the original intention of supporting modern web-scale operational databases and have begun to be used for analytics of the same data they operationalize. They store documents with encoded key-value pairs.	MongoDB, CouchDB, Cassandra, HBase, MemecacheDB, and others

There are hybrids that combine these categories in a variety of ways. Some vendors have incorporated columnar technology into an MPP RDBMS, or general purpose file systems and graph databases into an MPP RDBMS, or NoSQL engines into general purpose file systems or an MPP RDBMS.

Memory and Disk Choices

Engines can be configured with different storage types to achieve faster performance or to reduce costs when performance is less important:

- **Memory** — Systems can be configured with varying amounts of RAM. Some databases are designed to process all data in-memory
- **Disk** — There is a dizzying array of spinning disk options from High IO and High CPU to Low IO and Low CPU drives. Add to this solid-state drive storage, which does not spin and offers very fast retrieval as a result (for a price of course)

When looking at the available engines, it's important to separate the engines themselves from the way you could turbocharge them. Putting an engine in-memory is a separate issue from the type of engine it is (which could exist either in-memory or not, depending on the design choice, business need, and budget).

Design Pattern Choices

The engines just described can be used in support of various design patterns, including EDW, data lake, and discovery.

An EDW is a design pattern in which data is stored in a manner that is secure, cleansed, reliable, and easy to retrieve and manage. The schema for an EDW is defined in advance, referred to as ***schema on write***. Data modeling and the creation of entity relational diagrams are important parts of defining schemas and data structures.

A data lake is a methodology predicated on a massive data repository, enabled by low-cost technologies that improve the capture, refinement, and exploration of raw data within an enterprise. The data lake provides benefits such as:

- Cost effectively exploring datasets of unknown, underappreciated, or unrecognized value
- Improving the ETL process and the amount of information stored by offering a single source of raw, historical data
- Reducing the need for LOB-specific big data environments with lower costs and fewer analytical discrepancies
- Light, on the fly integration through the colocation of datasets

A data lake is a design pattern where data is collected in raw form. It can be read through multiple "lenses" depending on the use case, a convention referred to as ***schema on read***.

Discovery is another design pattern in common use today. Discovery systems typically support light ETL and data modeling to accelerate performance and ease of use, thereby making the discovery process iterative and democratic, while preserving much of the flexibility realized in a data lake approach.

While technologies and design patterns have historically had strongly correlated relationships, that is no longer the case as vendors have incorporated essential features to deliver on many key aspects of the above design patterns. For example, MPP RDBMS technology and the EDW design approach are often used interchangeably; however, some MPP RDBMS vendors have introduced the ability to store native data types and execute schema on read techniques. Similarly, some companies elect to implement defined schemas within Hadoop.

Economic Choices

Capabilities and the resulting revenue potential must also be evaluated with respect to costs, and costs should be evaluated across capital expenditures and operational expenses associated with development, usage, maintenance, support, and data center resources.

While extensive and detailed pro-forma financial calculators exist to arrive at TCO for big data projects, there are glaring realities that again reinforce the need for a logical data warehouse comprised of different components, including:

- **Hardware choices** — In-memory is significantly more costly than slow, dense spinning drives and there are many points between these extremes
- **Software choices** — Commercial software with mature features has higher acquisition costs than newer open source software products
- **Staffing choices** — Mature, proven software is often easier to configure and support than emerging software products, which require a higher skill level for developers and administrators
- **Data management choices** — Costs with an EDW are incurred primarily during development, whereas data management costs with a data lake are incurred primarily during consumption

The wisest course is for companies to pursue value, with the economics underlying the use of data and platforms dictating analytics use. Given this reality, and the range of choices and capabilities, no single platform is perfect. We need multiple platforms. The question is how to make exploiting multiple platforms easier.

No single platform is perfect. The question is how to make exploiting multiple platforms easier

A Complex Matrix of Options

Of course, the downside from all these options is the inevitable complexity that arises. Organizations are challenged to capture and interpret data that is spread across various analytics systems, each system handling different types of processing and data. Faced with these challenges, companies risk making mistakes from the past, where data is locked into silos or duplicated and moved in bulk between systems. The siloed approach limits analytics insights, and duplicating and moving data results in increased costs, latency issues, and different answers. It also places a huge burden on the IT organization to keep up with the business demands for data integration.

Further, many new big data technologies have primitive interfaces and languages that limit end-user adoption and connectivity with mature ANSI SQL-based applications.

The complexity is potentially overwhelming, and good intentions around deploying enhanced business capabilities can result in a disjointed and uncoordinated environment that fails to deliver on the full promise.

Data powers business, but the need to integrate results from many different engines bogs down operations. In many cases, analysts may not leverage all available data if the time and effort exceed its potential value. Companies are deriving less value from analytics than they should.

The need to integrate results bogs down operations. Companies are deriving less value from analytics than they should.

The Way Forward: Leverage Multiple Choices and Capabilities with Teradata QueryGrid™

A world of multiple analytics engines is inevitable. The challenge is how to embrace the benefits while minimizing the complexity. Analysts should focus on the story the data can tell, not where it resides. Paying analysts to be data movers, manually moving data in and out of various engines, is not the best use of resources.

Analysts should be able to find and access the data they need in a way that allows for ease of use, simplicity, and the ability to run whatever analytics make the most sense. Data should be tapped where it resides, significantly lowering the cost and time for analysis.

How Teradata QueryGrid™ Simplifies Analytics

Organizations are seeking the ability to scale the breadth and sophistication of their analytics to respond to the demands of business operations. The challenge is how to best orchestrate a wide variety of new analytics engines, file systems, storage techniques, procedural languages, and data types into one cohesive, interconnected, and complementary analytics architecture.

Data has varying degrees of business value, sometimes referred to as business value density. It is very important to decide where to house data based on its business value density along with the business need to access that data. Other considerations include the performance requirements for retrieving the data. When customers are waiting (on the phone, at the website, on an app), response time is critical.

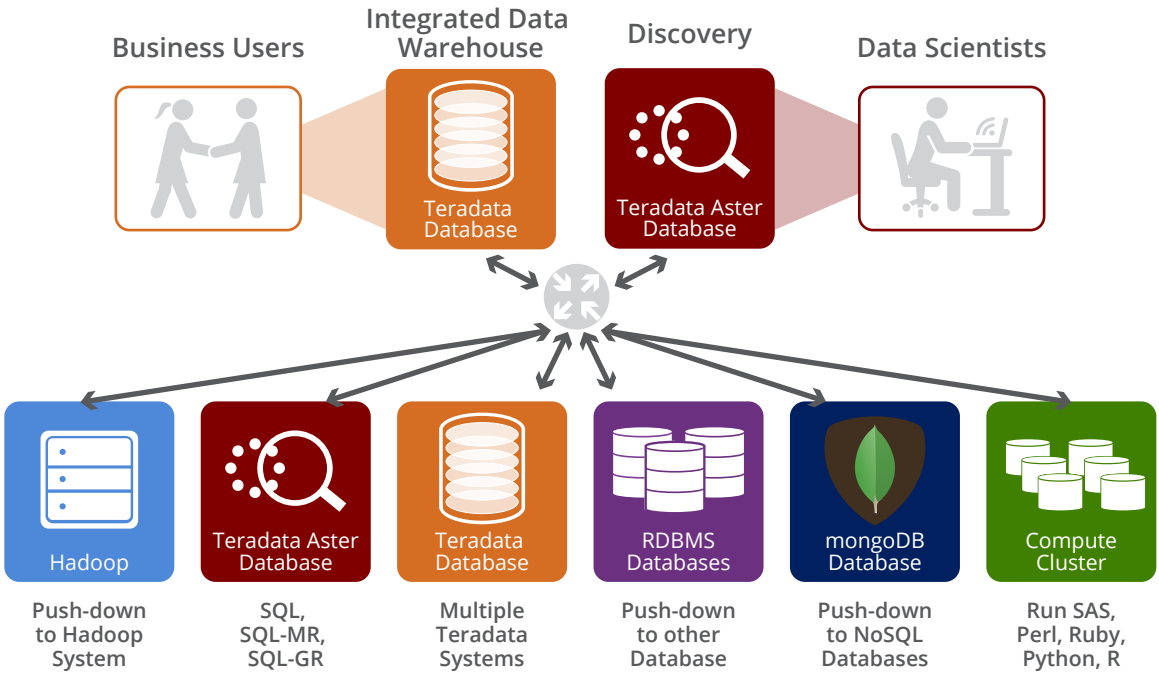
This paper does not cover the important question of where to store each type of data. This topic is covered in a paper called "Optimize the Business Value of All Your Enterprise Data."

In that paper, Teradata explained the rationale for creating a logical data warehouse using all the data stores available. The umbrella concept for this structure is the Teradata Unified Data Architecture™. The implied message of that paper is that as data is distilled and has higher business value density, it will find a home in the EDW. (In other words, once data is determined to have a high business value density, it should be modeled and ETL'd into the EDW so that it is more widely available.)

That paper did not address how data from multiple systems can be leveraged. What if analysts want to combine some data with high business value density stored in the data warehouse with some low business value density data stored in Hadoop?

Users tap into data where it resides by entering a single query that then reaches into the data and processing capabilities of each environment

Teradata addressed this gap by introducing a technology called Teradata QueryGrid™ that allows access to multiple engines through one interface. QueryGrid™ allows one query to reach into other data management and analytics platforms, then push down processing to those platforms, and return the smallest possible result set so the work of the comprehensive query can be completed. Users tap into data where it resides by entering a single query that then reaches into the data and processing capabilities of each environment. The query completes each of its steps in the best place to accomplish its task and returns the answer.



Teradata QueryGrid™ coordinates analytics across multiple engines

QueryGrid™ works by orchestrating a single business question into subqueries that have many sections, each of which represents the query for a different engine. Only the results, as opposed to the full dataset, of the subquery are moved into the EDW for final completion of the analytic.

Because QueryGrid™ handles the orchestration of the results, analytics can be conducted without IT intervention, empowering analysts to do what they do best. And, without the need to move data around constantly, IT is no longer a bottleneck to integrating or unifying data.

QueryGrid™ finally makes data work for businesses rather than having businesses work for their data. Each repository can do what it does best. Less data is duplicated because it persists where it is.

Teradata QueryGrid™ gives users seamless, self-service access to data and analytics processing across different systems from within a single Teradata Database or Aster Database query. Teradata QueryGrid™ uses analytics engines and file systems to concentrate their power on accessing and analyzing data without special tools or IT intervention. It minimizes data movement and duplication by processing data where it resides.

Driving the right offer

Retailers can gain new insight into their customers and drive incremental revenue. By streaming web events from the ecommerce store in Hadoop and creating a well-formed view of customers in the EDW complete with profitability, purchase history, and campaign history, a QueryGrid™ analytic can be constructed to execute a highly personalized and timely email offer to promote a product that a customer has browsed online but has yet to purchase, and to discount the offer at the optimal level based on past take rates and lifetime value.

Predictive maintenance

Manufacturers can capitalize on emerging sensor data to do proactive maintenance that lowers costs and increases customer satisfaction (and in some cases safety). By ingesting vehicle sensor data into a discovery engine, and creating a detailed supply chain model within the EDW, a QueryGrid™ analytic can invoke time series analysis to identify the cause of part failures. It can then link that to the antecedent sensor events that forewarned of the part failure and trace that back to the ultimate root cause in the supply chain.

Benefits of Teradata® QueryGrid™

What is QueryGrid™? It's a single interface that delivers the ultimate self-service experience for analysts by abstracting the complexity of the underlying robust ecosystem. Analysts get access to the data they want, when they want it, regardless of where it is. The benefits are profound:

- Run the right analytic on the right platform
- Automate and optimize work distribution through "push-down" processing across platforms
- Minimize data movement; process the data where it resides
- Minimize data duplication
- Transparently automate analytics processing and data movement between systems
- Access data and analytics easily using existing SQL skills and tools

Conclusion

Analytics is driving businesses of all types. It's not enough to have the data; companies have to use it to answer complex business questions. Companies must have a way to leverage all their data sources across their corporate landscape while reducing the friction in doing so. Otherwise, analysts cannot do their jobs.

Based on CITO Research's analysis, the Teradata UDA and QueryGrid™ are unique in the market and essential to empowering companies in realizing this vision. With this approach, data persists where it is and analysts can access all available data and analyze it on any platform through a single interface. As a result, the choices facing companies, from data storage systems to analytics platforms, are no longer overwhelming. Rather, companies come to view choices as opportunities, finally able to harness the unique capabilities of each engine without sacrificing processing time, cost, efficiency, or data reliability. With the Teradata UDA and QueryGrid™, businesses can embrace and benefit from all their data. It's not complicated anymore.

This paper was created by CITO Research and sponsored by TERADATA.

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com