Analytics in Action with Teradata QueryGrid™

Richard Hackathorn, Bolder Technology

Bolder Technology, Inc.

Table of Contents

This study examines business use cases for accessing Hadoop data from various Teradata platforms using Teradata QueryGrid™. Our interviews surfaced the business motivations and technical architecture for these use cases, while touching upon deeper issues concerning enterprise analytics at scale. Emerging is an approach enabled by federated workflows within an integrated information ecosystem among a fabric of interconnected purpose-built platforms.

Context

These are confusing times for IT management. Accepted practices and established technologies seem limited or even irrelevant in light of today's opportunities and challenges for incorporating big data and advanced analytics into enterprises systems. How should companies navigate this minefield to achieve business objectives? That is the context of this study.

Our starting point is the Teradata Unified Data Architecture™ (UDA)[1], which Teradata defines as "a design pattern that unifies multiple platforms within an analytic ecosystem." One of the key components of UDA is the Teradata QueryGrid™, as shown in Figure 1.

QueryGrid provides the access layer as a parallelized switching service between platforms so "business users do not care where the data is stored".[2] Chris Twogood of Teradata states succinctly "play it where it lies".[3] The 'it' to which he is referring is 'data' in various forms stored on various platforms.

The vision for QueryGrid is to orchestrate analytic processes across multiple platforms,[4] as a single unit of work based on SQL, minimizing data movement and leveraging unique processing capabilities among a fabric of platforms. QueryGrid also supports pushdown processing (remote process execution) so that remote platforms can perform specialized analytic processing. Hence, QueryGrid is extending Twogood's 'it' to process, as in "play the process where it lies".

Teradata states[5] that their investment in the product suite for Teradata QueryGrid will expand in several areas. First, Teradata is creating more pairs of parallelized interconnects, such as Teradata-to-Teradata, Teradata-to-Aster, and Teradata-to-MongoDB databases, plus more to come. Second, Teradata Labs will enhance Teradata QueryGrid with easier administration, performance enhancements, and workload optimization. New and innovative use cases for enterprise analytics will emerge over the coming years as this new functionality unfolds. This is an ambitious goal!

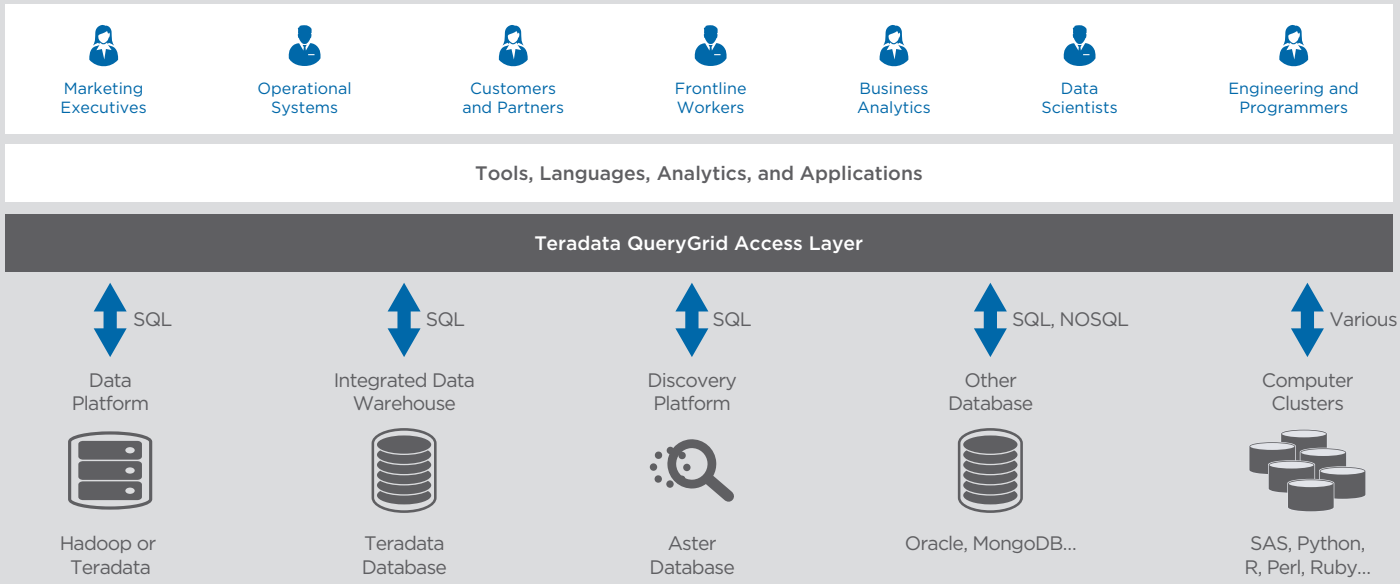

Figure 1. Teradata Unified Data Architecture.

Several years ago, Teradata and Hortonworks® developed the Teradata SQL-H™ and Teradata Aster SQL-H™ connectors,[6] which moves data from the Hadoop Distributed File System (HDFS) on the Hadoop platform to/from the Teradata or Teradata Aster systems. The innovation is these connectors parallelizes this data movement by mapping a processing unit on one platform to processing units on the other platform, thus achieving up to 100x throughput rates. Teradata has now merged these earlier connectors into the broader suite of QueryGrid connectors.

In this paper, we will use the following shorthand to reference the platforms for readability.

- Teradata platform = Teradata Integrated Data Warehouse™ with the Teradata Active Enterprise Data Warehouse™ (6xxx) or Teradata Data Warehouse Appliance™ (2xxx or 1xxx)
- Aster platform = Teradata Aster Discovery Platform™
- Hadoop platform = Teradata Appliance for Hadoop™ or customized Hadoop platform[7]

Experiences

This section describes actual customer experiences with the earlier QueryGrid connectors that access data on the Hadoop platform from the Teradata and Teradata Aster systems for reporting and analytics. The material is based on fifteen anonymous interviews with twelve knowledgeable professionals involved with nine different companies. The table below lists the companies, along with their general industry, platforms used and highlights of the use case.

Read this material for the business objectives, flow of data, and recurring issues. The next section will summarize the key themes across the fourteen QueryGrid use cases.

Vehicle Manufacturer

The company manufactures and operates large complex vehicular equipment with critical life and safety issues. The equipment is highly computerized with sensors monitoring every aspect of its operation. Sensor readings stream back to the company in batches throughout the day.

Industry	Platforms	Usage
Vehicle Manufacturer	Teradata, Hadoop	Bridging Cultures
Communications Provider	Teradata, Aster, Hadoop	Analytic Workflow
Financial Services	Teradata, Aster, Hadoop	Compliance/Security
Travel Services	Teradata, Hadoop	Parallel Streams
Computer Manufacturer	Teradata, Aster, Hadoop	Precision Views
Telecommunications	Teradata, Aster, Hadoop	Massive Discovery Lab
eCommerce Provider	Teradata, Hadoop	Website Search
Financial Systems Provider	Teradata, Aster, Hadoop	Travel versus WebEx
Electronics Manufacturer	Teradata, Aster, Hadoop	Process Control

Figure 2. Summary of Teradata QueryGrid Usage by Industry

A typical vehicle generates a few megabytes per day with several thousand data items, which is stored and maintained for about $1,000 per year. The sensor data is stored in Hadoop cluster, which enables low-cost access to the entire sensor history. The problem was "getting answers from the sensor data in a timely fashion" to improve the operation of that equipment.

The first use case was the daily maintenance of the equipment. By monitoring whether the sensor readings were within specified limits, the company could identify specific components for repair or replacement. However, the company realized that the potential was far greater, spanning many functions, such as failure predictions, accident investigations, long-term maintenance strategies, design testing, and operational efficiency. This was no longer a single-purpose application, but the beginning of an evolving infrastructure with impacts across their business.

Although the Hadoop cluster supported a variety of processing using Pig, Hive, and MapReduce, specialized skills required and complexity of coding limited the number of users accessing this platform. Further, the analyses required blending other reference data (such as equipment schedules, staffing) that was constantly changing and maintained on the Teradata data warehouse system.

The corporate IT staff and Hadoop developers needed to apply their distinct capabilities and collaborate to solve the larger issues of sensor analysis. These were two cultures with different skill sets, project timelines, management reporting lines, and degrees of urgency. The Hadoop people were "happy doing their MapReduce stuff" while the IT people "were happy doing SQL".

> Hadoop people were happy doing their MapReduce stuff while the IT people were happy doing SQL.

But some type of bridge was required between the infrastructures and the two cultures. One solution would be to extract sensor data from Hadoop and load it into the data warehouse. However, the data volumes and query diversity made this approach cumbersome.

About this time, Teradata released the QueryGrid connector that provided a bridge between the two cultures and resulted in a "more natural" work environment. Moreover, this bridge provided the basis for collaboration.

If the SQL queries produced datasets that were too large, data on Hadoop is reorganized to be more selective during a QueryGrid request.

The SQL OVERLAP and INTERSECT temporal extensions on the Teradata platform enabled a simple way to dissect complex time overlapping of sensor data, which is otherwise error prone. In addition, temporal compression achieved dramatic efficiency improvements, opening up new analysis opportunities for the company.

The second use case is detecting unnecessary maintenance. The company found a high percentage of "false positives," which caused vehicles to be serviced unnecessarily, thus increasing operational costs. The company is now better with their sensor analysis, reducing unnecessary maintenance costs.

Communications Provider

This company supplies communication services. The use case for QueryGrid is to improve customer retention by understanding customer behavior, such as when customers have negative experiences with their services.

Data volumes generated by all the customer touch points were too large to be managed cost effectively within the Teradata data warehouse, so the company moved this data to Hadoop. The company also acquired the Teradata Aster Discovery Platform to perform the required analytics. Their approach was to "store the data where it belongs and process the analytics where it belongs." QueryGrid connected the three platforms as a "marriage of relational data to nonrelational data."

The company started "bringing in new data sources that they never thought they would retain" by adding Hadoop storage. The data initially consisted of a few billion rows per day of company website logs, plus demographics and billing history. However, this data provided "a narrow view" of the customers. The company expanded data collection to incorporate many other web-based sources. Because customers are heavily researching and buying via smartphones, the data volume exploded by several orders of magnitude.

Event sequencing analysis (aka the golden path query) was difficult to perform with standard SQL that required many complex self-joins. The Teradata Aster SQL MapReduce using the Teradata Aster nPath™ function enabled the company to answer quickly event-sequencing questions like: *What steps does the customer do prior to using or buying our products? What is the last step that the customer performs before going away from our website? What are the steps leading up to a purchasing decision?* The company now has the ability to "analyze all these things at once, which is very important" for understanding customer behavior.

The Aster platform pulls weblog data from Hadoop and "marries" it with billing data from the Teradata data warehouse. The Teradata Aster SQL-MapReduce™ job on Aster initiates the QueryGrid transfer from Hadoop into a temporary or permanent table. It then performs the MapReduce processing on Aster. "The flexibility is great!" The company is able to paint the entire customer journey in using their product and services.

The company felt that QueryGrid will "get rid of a lot of busy work for their data scientists." The latest version allows MapReduce jobs to be pushed down to Hive on Hadoop so that only filtered results are transferred back to Aster. Since the data volumes are huge, this filtering is a huge advantage in performance.

The users of QueryGrid are business-oriented data scientists who prefer "higher level tools rather than getting into the weeds." These users are not "Silicon-Valley hard-core coders" but practical analysts who are "more interested in solving the business problem than creating elegant engineering projects." They will use "whatever tool to get the job done, taking the easy route. Time required to generate analytic results matters."

> Teradata QueryGrid will get rid of a lot of busy work for their data scientists.

The unique value of QueryGrid is both its ease-of-use and its efficiency. QueryGrid initiates a massively parallel connection. There is handshaking between the Aster worker nodes that are mapped to the Hadoop data nodes, resulting in hundreds of concurrent data streams between the platforms. Typically, this increases throughput by a hundred fold over a single stream.

This efficiency not only eliminates bottlenecks, but it changes the workflow of the analysts. Given enough time, any amount of data can be moved between the two platforms. But if an entire day is consumed with data movement, then the workflow is slower and disjointed. With QueryGrid, business analysts will ask more questions and get more answers per minute, resulting in more alternatives explored and better research of business problems.

Financial Services

The company provides consumer and commercial financial services. Over the years, the company's IT infrastructure has grown to over a petabyte of data managed among the Teradata, Aster, and Hadoop platforms.

The first use case is an application to monitor compliance of broker agents with client investment advice. The company must monitor all communications between brokers and their clients to ensure that the brokers are not promising investment performance. Because prior legal cases have resulted in regulatory fines, the company maintains hundreds of millions of dollars in reserves, which the company wants reduced. The old compliance system scanned broker emails for broker statements using a set of rules to detect improper activity. However, the analysis generated many false positives where an email is flagged as improper but is later judged to be proper. They employed hundreds of persons to evaluate all those flagged email messages.

The company did a proof-of-concept on whether machine learning could improve the email analysis accuracy of detecting improper emails. They decided to try the Aster text conditioning and analysis functions on the broker emails. Using Aster to perform the training/testing of a machine-learning model, the company improved bad email detection accuracy by 50%. The proof of concept has become a production application using Aster, Hadoop, and QueryGrid. Because of QueryGrid efficiency of parallel loads of Hadoop Hive data into Aster, the company is considering monitoring all corporate email, which amount to several million emails per day.

The second use case is weblog sessionization. Sessionization groups the sequence of customer web clicks into distinct customer visits in order to analyze preferences[8]. Over time, the company has used many software products to perform sessionization. However, they have found that the Aster platform does this function very well and could sessionize all their web channels, replacing the other software. This is important because weblog activity from their external (public) websites is separate from internal websites (where the customer signs into their account). A typical customer flips between public/private websites, often 70% of the time within the standard 30-minute session window. This causes the same customer to be treated as separate persons across separate websites, a problem that is fixed using Aster with QueryGrid.

Travel Reservations

The travel reservations company has three use cases for QueryGrid. The first two involve conversion funnels for the website and for the call center, while the third use case involves improving customer experiences on the website using A/B testing.

Customer journeys to the company's website can start at many locations, such as Google, meta-search, and sales affiliates. There are then further touches (clicks) from customers as they move through the website, such as categories within the site and within categories. These various paths may or may not be fruitful for the customer. Some customers will eventually book a travel service, and a transactional system of the company will capture the booking data.

As customers funnel through the company's website, the company's goal is to determine how customers convert from browsing to purchasing, which is referred to as the conversion funnel. To do so, the company must analyze two datasets. The first is behavior data from website logs and web analytic vendors. The second is the bookings captured by transactional systems.

The first use case for QueryGrid involves the conversion funnel using the website logs, so that the company can answer questions such as "Where should we spend marketing funds to improve website bookings?"

Web logs contain a significant amount of data that needs to be parsed to assess the unique customer journey for every customer visit. Hadoop is useful for this kind of processing. The bookings data is contained in ERP systems and transferred regularly into the Teradata data

warehouse. Therefore, QueryGrid brings the finalized customer journey data from Hadoop into Teradata to be combined with the bookings data, showing which customer journeys convert to revenue contributions to the business.

The second use case also involves the conversion funnel but using data from call centers in the form of IVR datasets, which are complex sequences of text and audio. The IVR data is also stored on the Hadoop platform. QueryGrid is used to combine the booking data with the IVR journeys to determine conversion contributions from customers.

The third use case for QueryGrid involves the support of A/B testing for improving website content design. The company is constantly testing new ideas for website content and style, such as listing travel alternatives to page layouts and special offers. Every day, the company assesses the impact of dozens of website changes observed by randomly selected customers.

> Experimenting with hundreds of data streams running over night, which they deemed to be unbelievably cool.

On the Hadoop platform, the A/B testing team tracks each website change within the click stream data. Throughout the day, QueryGrid is used to transfer booking data from the Teradata platform to the Hadoop platform. The team matches the two datasets to calculate metrics (like dollar volume) for each test.

The company realizes that they are just beginning to leverage the synergism of Teradata and Hadoop platforms. They are experimenting with hundreds of QueryGrid data streams running over night, which they deemed to be "unbelievably cool." Customer behavioral data from various sources is growing rapidly on the Hadoop platform. The company is not trying to integrate data within Hadoop, but they are experimenting with evolving that data into productive business applications.

The company recommends being "careful and thoughtful" about synchronizing data movement among platforms and optimizing processing where the data resides, thereby minimizing the data movement. However, users should access data through the platform that they prefer, striking a balance between the value propositions of data storage to analytic processing among the various platforms.

Computer Manufacturer

The company manufactures a variety of computer systems. Their use case for QueryGrid was to understand the customer journey through their website with a 360-view of buyer behavior. The benefits are primarily in lead generation, which are used by the sales organization. The company operates Hadoop, Aster, and Teradata platforms, all of which are InfiniBand™ connected for performance.

The weblog clickstream data is stored on the Hadoop platform. Customer data is collected via SalesForce.com in the Teradata platform. The Aster platform collects the relevant data from these two platforms, analyzes and shares the results using the Teradata Aster Lens™ visualizations, along with Excel downloads for final end users. The company used this arrangement because it "was faster than moving all the data to the Teradata platform."

The company uses QueryGrid via "precision" SQL base-views that hide the HDFS configuration to protect the data since Hadoop lacks adequate security. Upon the base-views, transform-views perform the required data transformations to feed the Aster analytics.

The company has compiled roughly 50TB of compressed data covering customer online activities over two years. They are expanding its use of Aster analytics, particularly with weblog sessionization, Aster nPath event tables, and eventually graph analysis.

The company noted their entire IT infrastructure has been centered on relational database using SQL for decades. "We can plug-and-play using traditional SQL," avoiding learning the skillsets of Java and Python. Their analytic team consisting of SQL developers is offshore. There are a "handful of Hadoop newbies familiar with R and its statistical functions." However, most users are directly using Aster, along with two data scientists who are "hacking away on Aster."

The coolest part of using QueryGrid is "Being able to have a big data lake without messing with the Hadoop zoo." They are retaining the data in Hadoop and pulling selective data into the Teradata platform using traditional SQL tools."

Telecommunications

The company is a telecommunication firm that is focusing on a complete 360-degree view of their customers. The company has two related use cases for the QueryGrid connector: customer churn and customer satisfaction index.

In both use cases, the Aster Discovery Platform is used for analytics. QueryGrid is "the quickest and easiest way to "peek at the data every day". Every day, production batches using SQL table operators refresh tables on Aster from the raw data on Hadoop. In addition, several power users submit ad hoc queries to pull, refine, and play with data for new analytic applications, which eventually are put into production if proven useful. Results generated on the Aster platform are shared among analysts using Tableau, Aster Lens, and Excel. The company plans to push results to Teradata for integration with enterprise data and Business Objects.

> The level of effort is really low once you understand how the connector works.

The first use case focuses on reducing customer churn switching to competitors, which includes improving the network quality of cellular signals that may cause churn. This effort requires detailed data gathered from various external systems and stored on the Hadoop platform. Currently Hadoop acts mainly as a large HDFS file store, with little MapReduce processing.

The second use case is a customer satisfaction dashboard for use in their call centers. The purpose is to improve customer retention and up-sell options. This use case is interesting because it is compiling a great variety of data from more data sources.

The QueryGrid connector pulls about twenty data sources into Aster from both the Hadoop and Teradata platforms, refreshing a hundred million rows of data. The company notes "performance has been good."

The business value of QueryGrid is its ability to ingest data into Aster Discovery Platform for analysis. The company notes, "The level of effort is really low once you understand how QueryGrid works."

Since Aster acts as "a massive analytic sandbox," the company is able to perform rapid prototyping of new applications with new datasets. The Teradata Database platform is "locked down," but it contains "the crown jewels data" needed for new application development. In contrast, the Hadoop platform can "ingest all that messy new data." The Aster platform sits in the middle, using QueryGrid to fetch data from both the Teradata Database platform and the Hadoop platform. Projects like the second use case previously "took 3-6 months, but now similar projects are taking 2-3 days" to ingest and analyze the data, all with greater functionality and stability. In addition, the "crown jewels" hosted on Teradata Database platform were protected from the discovery activities on the Aster platform.

The company has high interest in future versions of QueryGrid, but many at the company are "still wrapping their heads around" this technology. Several groups, including statisticians, are now comfortable with the Aster nPath function and are eager to use R to model their data.

eCommerce

The eCommerce company focuses on improving the effectiveness of online customers to find and purchase products matching their needs. The Teradata database system supports their enterprise data warehouse (EDW), which encompassed most of their business functions, while a separate appliance-based Hadoop platform supports behavioral analytics.

> Users are allowed to do the things they are best at on their preferred platform.

The company was an early adopter of the Hadoop platform, which was initially for "skunkworks projects focused on unique hard problems that were considered unsolvable through existing technology solutions." Over a few years, the company gained confidence in the Hadoop platform and blended it with their mainstream IT infrastructure.

The early applications that used the Hadoop platform often required most of its data from the EDW. Hence, an early requirement was to move data in both directions between Hadoop and Teradata platforms.[9] This is

different from typical Hadoop adopters who compile massive datasets from website click-streams and generate small result sets to be transferred to their EDW.

One application took text data from the EDW, performed text mining on Hadoop, and returned results to an external search provider. Later, other applications transformed data and moved results back to Teradata. The motivation for these was mainly for cost savings by reducing the Teradata workload. More recently, there is blending of the platforms so that "users are allowed to do the things they are best at" on their preferred platform.

Financial Systems Manufacturer

The company manufactures, installs and services financial systems worldwide. To perform these activities, employees are required to visit customers, service equipment, and coordinate with colleagues. For many years, the company has tried to reduce these travel expenses (along with unproductive time on airplanes) with telepresence video conferencing and WebEx desktop sharing. They showed that, in many situations, travel expenses could be reduced with virtual meetings, without affecting business performance.

The use case for QueryGrid is to identify who-interacts-with-whom within the company, correlate with travel events, and generate reports about actual/potential cost savings from virtual meetings. These reports were referred as 'guilt trip reports'.

A Hadoop platform collected detail call/message data at the IP phone-address level. The Aster platform did the initial analysis using the Aster analytics package (especially the nPath function) plus Tableau for visualization. Once the analytics team identified interesting results, management disseminated the 'guilt trip' reports on the Teradata platform via Business Objects reports. Hence, QueryGrid was utilized twice to extract data from Hadoop, initially from Aster and then from Teradata. In both situations, the use of QueryGrid could be hidden from the users of Tableau and Business Objects with SQL views so that Hadoop appears as a local database table.

Finally, the company admitted that this was an ad hoc application initiated "just to see if we could do it"! This is an example of the thousands of new analytic-based applications that are possible with current analytic tools "...if we nudge ourselves to think creatively."

Electronics Manufacturer

The electronics manufacturing company focuses on detecting potential failures in their manufacturing process. Based on hundreds of second-by-second sensor readings, they want to predict when product quality will exceed its limits, so that they could quickly stop the assembly line and correct operating parameters on manufacturing equipment.

The company was collecting sensor data but only a selected subset. They felt that, if they could apply analytics to all the data, their failure detection would improve. The company has been prototyping a Hadoop platform to collect all of the sensor data (over ten gigabytes per hour per assembly line) within HDFS. Their Hadoop prototyping proved that they could collect all the sensor data without problems. However, it was difficult to extract the data out of the Hadoop platform within their requirement for fast response times.

After proof-of-concept testing with several vendors, they decided to acquire a Teradata Data Warehouse Appliance and use QueryGrid to extract the data from their Hadoop platform over a 10 Gb network. The Teradata platform will be used to analyze the sensor data and to generate reports and dashboards to a wide spectrum of business users throughout the company. Currently they are assembling and testing the systems with full operation in 2015.

Perceptions

This section summarizes the use-case experiences with QueryGrid as a series of insightful perceptions by the companies. The table below lists the use cases, along with the QueryGrid connections among platforms and the maturity for the application. Daily use implies manual activities using QueryGrid, while production indicates that QueryGrid is embedded in automated processing.

Under QueryGrid connections, the bolded platform indicates the local platform that initiates a connection to the remote platform. Note that, in many use cases, the Aster platform plays a 'linking' role between the Hadoop and Teradata platforms.

Industry	Use Case	QueryGrid Connection	Maturity
Vehicle Manufacturer	1. Indentify components needing maintenance	Teradata-Hadoop	Daily Use
	2. Detecting unnecessary maintenance	Teradata-Hadoop	Prototype
Communications	3. Improving customer retention	Teradata-Aster-Hadoop	Daily Use
Financial Services	4. Monitoring brokerage compliance	Teradata-Aster-Hadoop	Production
	5. Processing weblog sessionization	Teradata-Aster-Hadoop	Prototype
Travel Services	6. Conversion funnel using website logs	Teradata-Hadoop	Daily Use
	7. Conversion funnel using IVR call center logs	Teradata-Hadoop	Prototype
	8. Improving website design with A/B testing	Teradata-Hadoop	Daily Use
Computer Manufacturer	9. Generating leads from customer journey	Teradata-Aster-Hadoop	Daily Use
Telecommunications	10. Reducing customer churn	Teradata-Aster-Hadoop	Daily Use
	11. Customer satisfaction dashboard at call centers	Teradata-Aster-Hadoop	Prototype
eCommerce	12. Improving website search for online customers	Teradata-Hadoop	Production
Financial Systems Manufacturer	13. Reducing travel costs	Aster-Hadoop & Teradata-Hadoop	Production
Electronic Manufacturer	14. Monitoring process quality control	Teradata-Hadoop	Prototype

Figure 3. List of Teradata QueryGrid Use Cases

Bridge Between Cultures

A recurring theme behind the QueryGrid use cases is the cultural bridge enabling two technical cultures to collaborate within the same information ecosystem. This theme is described as a cultural bridge resulting in a more natural work environment and as a marriage of relational data to nonrelational data. The Computer Manufacturer notes that they are able to have data in a big data lake without dealing with the Hadoop zoo but instead using traditional SQL tools.

The benefits cited for this cultural bridge are: technical people are happier, busy work for data scientists is reduced, learning new skillsets is avoided, and employee skills are leveraged—all of which stimulates innovations for the company.

Companies Referenced: Vehicle Manufacturer, Communications, Travel Services, Computer Manufacturer, eCommerce

Data Placement

The companies realized that the placement of data and its processing is an important configuration issue for their analytic infrastructure. The Communication Company summarizes the issue as simply storing the data where it

belongs and processing the analytics where it belongs. ...assuming that there is a clear choice about where it belongs! The Travel Services Company recommends being careful and thoughtful about synchronizing data movement among platforms to optimize processing where the data resides, thereby minimizing the data movement. However, the company also wants users to access data through the platform that they prefer, while striking a balance between the value propositions of data storage to processing among the various platforms.

For example, the Computer Manufacturer collects data on the Aster platform from both the Hadoop and Teradata platforms because this arrangement was faster than moving all the data to the Teradata platform. Likewise, the Telecommunication Company pulls about twenty data sources into the Aster platform from both the Hadoop and Teradata platforms, enabling Aster to act as a massive discovery lab.

Companies Referenced: Communications, Travel Services, Computer Manufacturer, Telecommunications, eCommerce, Financial Systems Manufacturer, Electronic Manufacturer

Data Marriages

Companies realize business value when the newer data from web or sensor sources is married with older reference data on customers, purchases, and the like. Normally residing in the data warehouse, reference data indicates what is important within the newer data, extending its semantics. Value is increased when results are married to the proper analytics. Value is further increased when results are married with reporting and dissemination tools. This set of marriages drives the justification for data movement among the platforms.

For example, the Travel Service Company needs the booking data from the Teradata platform to complete their conversion funnel and A/B testing analyses on the Hadoop platform. The Financial Systems Manufacturer uses Business Objects on the Teradata platform to disseminate results from the Aster Platform.

Companies Referenced: Communications, Travel Services, Financial Systems Manufacturer

Lot of Messy Data

Another theme was the motivation behind the adoption of the Hadoop platform. Instead of the sexy MapReduce and other higher-level functions, the critical requirement is quick and efficient data storage using the Hadoop Distributed File System (HDFS) with basic Hive. Messy data includes weblogs, sensor data, text, and social media.

Several companies were early adopters of the Hadoop platform for data discovery by their data scientists. Later, those companies expanded the Hadoop platform and incorporated it into the enterprise infrastructure driven by the need to access that messy data to support new business applications. Once the Hadoop platform proved its business value, the Communication Company remarked that they started bringing new data sources that they never thought they would retain.

Companies Referenced: Communications, Computer Manufacturer

Sequence Those Events

For many, the killer analytic app is the Aster nPath function, which discovers the event sequence that precede a significant business event, such as a customer switching to a competitor. The Communication Company notes that this event sequencing analysis (aka the golden path query) was difficult to perform with typical SQL that required many complex joins. This event sequencing analysis is now being applied to other business situations than customer churn.

Companies Referenced: Communications, Computer Manufacturer, Telecommunications, Financial Systems Manufacturer

Parallelizing Data Streams

In the early days of columnar database systems, query processing of complex SQL had huge (many orders of magnitude) performance improvements, thus enabling innovations. The same performance explosion is occurring with data movement within tightly coupled clusters of nodes, thus enabling innovations that leverage the Hadoop platform. For example, the Travel Services Company experimented with hundreds of parallelized data streams running over night, which they deemed to be unbelievably cool.

The benefit is not only eliminating bottlenecks but also changing the workflow of the analysts. Per unit of time, analysts will ask more questions and get more answers, thus resulting in more alternatives explored and better validation of business solutions. Time to analytic result matters.

Companies Referenced: Communications, Travel Services

SQL Views

The SQL views is cited several times as a way of simplifying usage and insuring security for QueryGrid. For example, the Financial Systems Manufacturer uses SQL views to hide QueryGrid from the users of Tableau and Business Objects so that the Hadoop data appears as a normal local database table. Further, the Computer Manufacturer uses *precision* SQL base-views that hide the HDFS configuration to protect data because Hadoop lacks adequate security.

Companies Referenced: Computer Manufacturer, Financial Systems Manufacturer

Reflections

In this last section, let us take a step back and reflect on these QueryGrid use cases. What do these experiences tell us about the future of enterprise analytics?

The IT industry is in the beginning stages of redefining the enterprise data warehouse. Emerging is an approach to support enterprise analytics at scale, as enabled by federated[10] workflows within an integrated information ecosystem[11] among a fabric[12] of interconnected purpose-built[13] platforms.

This last sentence is densely packed, so let us unpack it by recognizing the following points:

- A company operates within a complex global ecosystem …financially, politically, culturally and so on. The information ecosystem for that company must reflect the nature and structure of that global ecosystem.
- Analytics must become an integral component of the IT architecture, embracing a perpetual cycle of discovery and operationalization, which will profoundly challenge current data governance practices.
- The marriage of newer data from web and sensor sources with older data from the data warehouse is the major generator of business value.
- Smart workflow management is required for optimizing placement of both data and process across the information ecosystem. Data + Process are two sides of the same coin. Optimizing one will degrade the other.
- Heavy workflow requires parallelized dataflow utilizing high capacity interconnect fabric technology, such as InfiniBand® with Teradata BYNET®.
- The various technical cultures must build bridges of collaboration so that companies can leverage a variety of tool and skill sets. Best practices in analytics indicates that ensembles of models using various techniques and packages (R, Python scikit-learn and others) produce the best results.

Finally, the most difficult challenges for realizing enterprise analytics at scale do not involve technology, which is abundant. The political, cultural, and ethical aspects are the keys to success for any company striving toward enterprise analytics at scale. Combined with savvy application of technology, executives must focus their energies on these other aspects.

Endnotes

1 Teradata Unified Data Architecture, www.teradata.com/products-and-services/unified-data-architecture/

2 Teradata QueryGrid, http://www.teradata.com/Teradata-QueryGrid/

3 Chris Twogood, Teradata briefing on MapR partnership, November 17, 2014.

4 Imad Birouty, Harmonious Orchestration, Teradata Magazine, Q2 2014, http://www.teradatamagazine.com/v14n02/Tech2Tech/Harmonious-Orchestration/

5 www.teradata.com/News-Releases/2014/Data-Fabric-Enabled-by-Teradata-QueryGrid--Orchestrates-the-Analytical-Ecosystem/

6 http://www.teradata.com/News-Releases/2013/Teradata-Delivers-Industrys-First-Flexible-Comprehensive-Hadoop-Portfolio/

7 Hadoop also refers to a large dynamic technology community that is difficult to summarize. The Wikipedia entry (http://en.wikipedia.org/wiki/Apache_Hadoop) is good for an overview of its components. The Apache Hadoop site (http://hadoop.apache.org/) is good for the core open-source projects, such as Hive, Spark, etc. Most companies use specific code distributions (with unique enhancements) from Hortonworks, Cloudera, MapR and others.

8 For background on click-stream sessionization, see http://en.wikipedia.org/wiki/Sessionization

9 Because of the early requirement for Teradata-Hadoop data transfer, the eCommerce company did not use the Teradata QueryGrid connectors, which was not available at that time. Instead, they developed their own version, which was similar in function.

10 Federated is defined as "united into a single organism where components retain some degree of autonomy".

11 Ecosystem is defined as "community of interacting organisms with a common set of purposes".

12 Fabric is defined as "a cloth-like structure that weaves together different resources to work as a single entity".

13 Purpose-built is defined as "being designed and created for a specific purpose (objective or use), as opposed to being created with certain technology, architecture, or methodology".

About the Methodology

The methodology for this study is to listen carefully to pioneering companies in big data analytics and to report accurately their perceptions. The intent is to contribute to professional education—to share the experiences and best practices with other IT professionals so that we can mature as an industry, amid escalating business challenges and rapidly evolving technology.

The primary author is Richard Hackathorn of Bolder Technology, who appreciates the insights shared by a spectrum of IT professionals. Dan Graham of Teradata deserves special credit for his constructive criticism that resulted in substantive improvements to this study.

Finally, a sincere appreciation to Teradata Corporation for their sponsorship in conducting this study and for permitting open and independent access to their community.

About Bolder Technology

Bolder Technology Inc. is a twenty-year-old consultancy focused on Business Intelligence and Data Warehousing. The founder and president is Dr. Richard Hackathorn, who has more than thirty years of experience in the Information Technology industry as a well-known industry analyst, technology innovator, and international educator. He has pioneered many innovations in database management, decision support, client-server computing, database connectivity, and data warehousing.

Richard was a member of Codd & Date Associates and Database Associates, early pioneers in relational database management systems. In 1982, he founded MicroDecisionware Inc. (MDI), one of the first vendors of database connectivity products, growing the company to 180 employees. Sybase, now part of SAP, acquired MDI in 1994. He is a member of the Boulder BI Brain Trust (BBBT). He has written three books and has taught at the Wharton School and the University of Colorado. He received his degrees from the California Institute of Technology and the University of California, Irvine.

About Our Sponsors

Teradata helps companies get more value from data than any other company. Our big data analytic solutions, integrated marketing applications, and team of experts can help your company gain a sustainable competitive advantage with data.

For 20 years, Teradata Data Warehouse and Appliance solutions have seamlessly integrated NetApp E-Series storage for unparalleled performance, unlimited scalability, and nonstop availability. These finely tuned joint solutions are perfectly optimized to help accelerate business insight and drive your competitive advantage.

01.15 EB7044

BOLDER Technology, Inc.