



Data Exploration and Discovery: A New Approach to Analytics

Colin White, BI Research
October 2013

Sponsored by:

TERADATA.

TABLE OF CONTENTS

Introduction	1
Why Data Discovery?	1
Data Discovery and Big Data	2
Data Discovery versus Data Science	4
Data Discovery Use Cases	5
The Data Discovery Development Process	6
Technology Requirements	7
Getting Started on a Data Discovery Project	8
Vendor Example: Teradata Aster Discovery Platform	9
Teradata Aster Product Architecture	10
Data Discovery Capabilities	11
Teradata Strategy	12

INTRODUCTION

Data exploration and discovery enables information workers, such as data scientists, business analysts and power users, to blend and explore different types of data and analytics to look for ways of improving business processes and to discover new business opportunities.¹ It allows these users to investigate different types of data and analytics before committing any particular solution to a production business intelligence and decision-making environment.

Data discovery offers a new and improved way of designing and building analytic solutions. It provides faster time to value because in a data discovery system the data does not have to be integrated into a data warehouse before it can be analyzed. It also enables organizations to exploit the benefits of new big data technologies and extend traditional business decision making with solutions that increase the use and business value of analytics throughout the enterprise.

This paper discusses the role of data discovery in decision making, its benefits to the business and examples of use cases. It also looks at technology requirements and how to get started on a data discovery project. Lastly, it reviews the Teradata® Aster® (the sponsor of this paper) Discovery Platform as an example of how vendors are rapidly extending their product portfolios to provide the business value data discovery brings to organizations.

WHY DATA DISCOVERY?

Customer surveys and case studies clearly demonstrate that analytics are no longer just nice to have, but are essential in helping companies reduce costs, grow revenues, improve customer retention and satisfaction, and increase overall operational efficiency and productivity.

The growing use of business analytics by organizations is fueling a rapidly expanding vendor marketplace for analytic tools and applications. The analytics industry is moving beyond basic reporting and online analytic processing (OLAP) to provide a range of advanced capabilities for statistical, text and graph analysis, forecasting, predictive modeling and analysis, and optimization (see Figure 1). These capabilities (if used wisely) can add significant business value, but also often come at the cost of increased complexity. This is why vendors are also focusing on making their analytic solutions easier to use, because otherwise the skills requirements to use these advanced capabilities will become a barrier to their successful deployment and growth.

¹ For simplicity, this paper shortens the term *data exploration and discovery* to *data discovery*. Some industry experts and vendors further shorten *data exploration and discovery* to simply *discovery*. BI Research also uses the term *investigative computing* to describe *data exploration and discovery*.

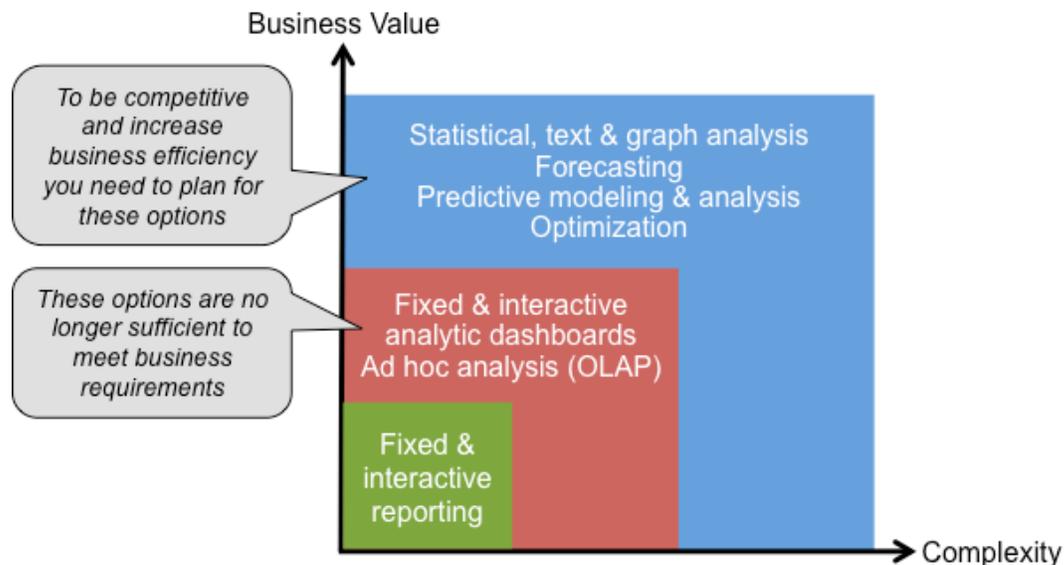


Figure 1. The three stages of BI evolution

As the analytics industry moves towards improving the power, business benefits and usability of analytics, the main model for building and deploying analytic applications has remained largely the same – data is extracted from operational systems, modeled and transformed, and then loaded into an enterprise data warehouse (EDW) for analysis. While this *classic* approach to creating analytics is essential to give users an integrated, consistent and governed view of data, the need to integrate data into an EDW before it can be analyzed can become a bottleneck to exploring new types and sources of data. Data discovery helps alleviate this bottleneck (see Figure 2) by allowing data to be investigated and analyzed outside of the EDW environment. This enables rapid and iterative data analysis and provides faster time to value when compared with classic approaches to analytics development.

It is important to emphasize that data discovery does not replace the EDW. Instead, it allows organizations to experiment with new data sources and new analytic techniques and models to improve and extend the EDW and the existing decision-making environment. The EDW can in fact be both a data source and a data target for a data discovery system.

DATA DISCOVERY AND BIG DATA

There are many different definitions for *big data*, but broadly speaking the benefits of big data fall into three main areas:

1. **Access to new sources of data** for improved and more informed decision making.
2. **New and enhanced analytic techniques and technologies** that increase the analytic power of existing decision making solutions.

3. **Improved analytic performance** enabling more sophisticated and new solutions to be built that were not possible before because of price/performance and/or technology limitations.

All of these benefits can be leveraged in a data discovery environment. In fact, it is the advent of big data, and the need to support many different big data scenarios that led to the development of many data discovery products.

Issues With Classic Analytics	Data Discovery Benefit
IT backlog and lack of resources to extend the EDW to support new business requirements	Data can be explored and analyzed outside of the EDW environment before it is put into production use
Costs of supporting increasing data volumes and new types of data	Data can be filtered and transformed before it is loaded into the EDW
Lack of flexibility in the EDW data model to support constantly changing business requirements	Data discovery supports dynamic <i>schema on read</i> approach, which reduces the need for detailed up-front schema design
Need to have data quality and governance processes in place before users can access EDW data	The investigative nature of data discovery has lower data quality and governance requirements
Growing use of personal data marts to overcome IT barriers and the performance overheads of <i>ad hoc</i> processing	The flexibility and performance of data discovery encourages shared use of data and analytics

Figure 2: The benefits of data discovery

The role of big data is different for each organization and each data discovery project. These differences can best be illustrated by examining two quite different big data scenarios.

Analyzing customer behavior across multiple channels. The business need here is the ability to blend existing customer information with a variety of new data sources (high-volume web store data and/or social media content, for example) to identify ways of improving web site design and operation, tuning marketing campaigns, optimizing offers to high-value customers and so forth.

Analyzing genome data. Unlike the previous example, this scenario does not involve multiple data sources or high data volumes – the data for a human genome easily fits onto a DVD. Instead, it involves highly complex and iterative analytic processing, and so sound analytic capabilities and performance are the main requirements here.

These two scenarios have quite different business and technology requirements, but as in all big data projects, the project is not just about the data, but also about the analytics and business value that can be derived from that data. There are, therefore, many variables that need to be considered when evaluating the use of big data for any given data discovery project and in selecting the most appropriate data discovery platform and tools. The main ones are:

- Amount of data to be managed (data volume)
- Rate of data generation or change (data velocity)
- Types of data to be managed (data variety)
- Number of data sources and data relationships, and the quality of the data (data complexity)
- Types and complexity of the analytic processing (workload complexity)
- Data currency and analytic application response time requirements (workload agility)
- Makeup of the total analytic workload (workload mix)

DATA DISCOVERY VERSUS DATA SCIENCE

People often perceive data discovery as simply another term for data science. Although the two fields are related and overlap (see Figure 3), there are some important differences.

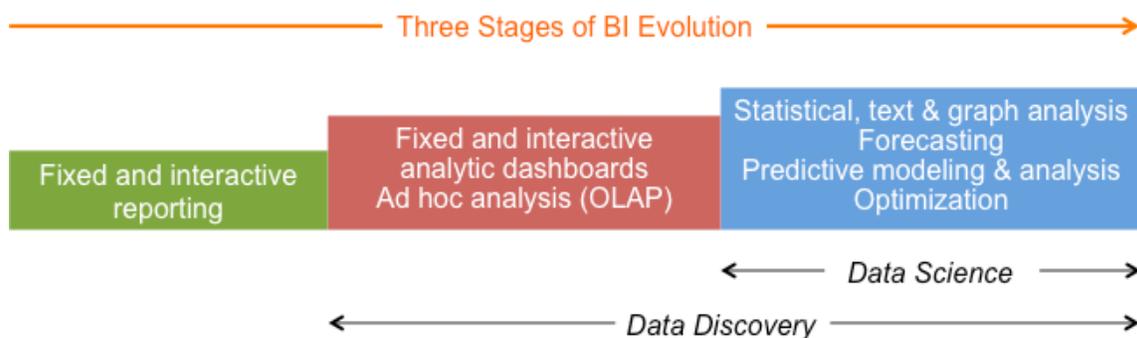


Figure 3. Data Science versus Data Discovery

Data science is an evolution of data mining, statistics and machine learning. In addition to supporting sophisticated analytic techniques and technologies, data science also places strong emphasis on business and IT collaboration, and on identifying clear business requirements. The objective of data science is to move advanced analytic techniques from a lab environment to being used in mainstream day-to-day business processes.

One of the main barriers to the successful deployment of data science projects in organizations is the lack of skilled staff. Data science requires knowledge of data

engineering and analysis, business analytics, advanced statistics, predictive modeling and analysis, and expertise in the business area being addressed. This is a tall order, and although universities and vendors are investing in data science education, the only practical solution to the skills problem is to build a data science team consisting of data analysts, business analysts, and experts in statistics and predictive modeling and analysis.

A data discovery system includes data science technologies and supports data scientists, but it also provides tools that can be used by business analysts and power users. These tools provide traditional OLAP capabilities as well as easier-to-use data science capabilities that are designed for less experienced users.

Some data discovery products also allow the user to combine multiple styles of analysis for richer results, for example, statistical analysis (traditional customer data, for example) with text analysis (social media data) and graph analysis (social networking data). The idea here is to provide business analysts and power users with a range of tools they can use to explore data and identify business trends and patterns. Patterns of interest can then be used to optimize business processes – they can also be given to data scientists for more detailed analysis. This approach reduces the load on the data scientist since business analysts and power users are doing more of the upfront work. It also encourages improved communication between different types of information workers.

DATA DISCOVERY USE CASES

About half of all data discovery projects at present are related to customer-facing business processes. An example here is the analysis of customer behavior. Traditional OLAP tools can be used to report on the types of products customers are buying, but they often don't help, for example, in understanding why customers buy certain products, or why they place products in a web shopping basket, but don't actually purchase them. OLAP tools also don't help in bringing other aspects of customer buying behavior (such as customer sentiment and satisfaction) into the analysis process.

Data discovery helps analyze customer behavior and sentiment by bringing together customer data from a variety of different sources for processing by a variety of more sophisticated analytic capabilities than those offered by OLAP. Initially these tools may be used for descriptive analytics, i.e., analyzing past customer behavior and optimizing existing marketing campaigns, but as data scientists become more involved in the data discovery process, the patterns identified by business analysts can be used to build predictive models, for example, to predict churn by telecommunications or banking customers. This approach enables organizations to move toward the use of more advanced analytic techniques in a phased and evolutionary manner.

Additional examples of the use of data discovery in customer-facing applications include sales funnel analysis, customer micro-segmentation, advertising effectiveness, social network analysis, warranty analysis and fraud detection.

Beyond customer-facing solutions, other areas being addressed by data discovery include finance (anti-money laundering and risk management) and operational efficiency

(identifying waste, reducing costs, supplier quality and performance, and optimizing supply chain operations).

Significant longer term potential for data discovery lies in analyzing sensor data. These latter solutions will be industry specific. Examples of industries that are deploying or evaluating data discovery platforms for analyzing sensor data include security, utilities, oil and gas exploration, healthcare, telecommunications and transportation.

THE DATA DISCOVERY DEVELOPMENT PROCESS

Data discovery represents a significant change to the way organizations design and build analytic solutions. This can best be explained using Figure 4, which illustrates the data flow in a data discovery system.

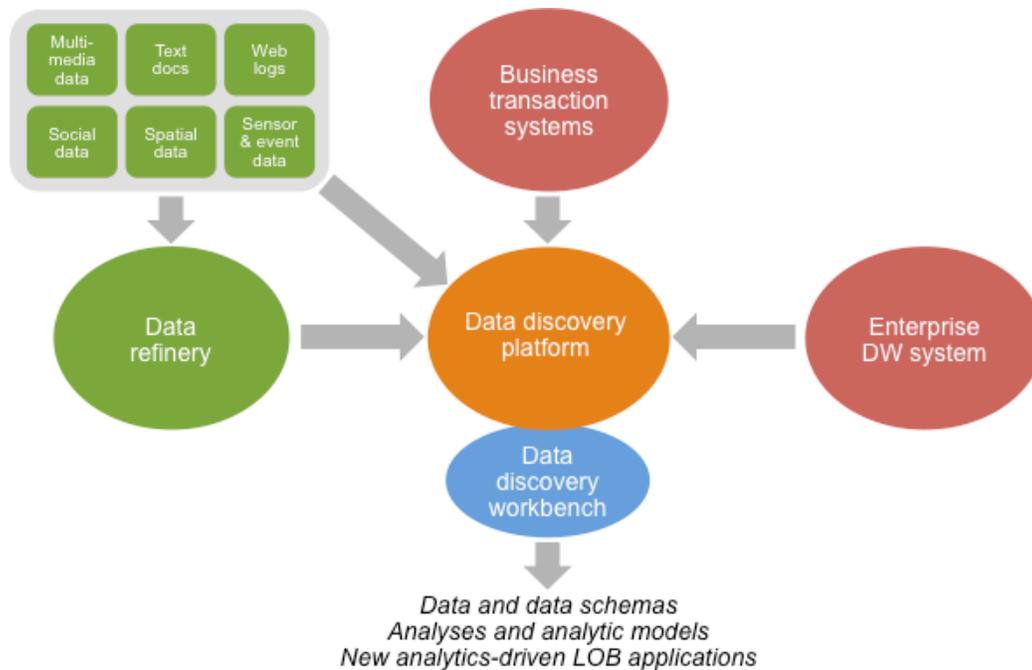


Figure 4. Data flow in a data discovery system

As shown in Figure 4, a data discovery platform gathers data from a variety of sources, blends it together, and then analyzes it. Data may come from structured data sources such as a data warehouse and business transaction systems, from multi-structured data sources like document management systems and web-based platforms, or from sensors on intelligent hardware devices. The output from the data discovery process may include data and data schemas for deployment in a data warehouse, analyses and analytic models for use in a production business intelligence system or automated business process, or a new analytics-driven line-of-business (LOB) application.

When large volumes of raw multi-structured data are involved, the source data may be pre-processed by a *data refinery* prior to it being used by a data discovery system. Many organizations are beginning to build data refineries running on systems such as Hadoop as a cost-effective way of managing and transforming large volumes of raw data.

One of the main differences in a data discovery system, as compared with a traditional data warehousing workflow, is that the information worker can blend, explore, analyze and visualize data in different ways without the need for rigid pre-defined data schemas and data integration workflows. This flexibility is provided by a *data discovery workbench*, which includes a set of tools that allow information workers to dynamically build data schemas and blend data together as it is read from various source systems.

The data discovery workbench also provides a variety of different tools for analyzing and visualizing the blended data. It may, for example, include tools for OLAP, statistical and text analysis, forecasting, predictive modeling and analysis, and/or optimization. The actual tools used will depend on the skills of the information worker – power users and business analysts, for example, are typically less skilled than data scientists and therefore require less sophisticated and easier to use tools. The tools may be supplied with a data discovery platform (such as the Aster Discovery Platform) or provided by a third-party supplier (such as SAS, Tableau or Tibco). Many organizations are likely to use a combination tools from both data discovery platform and third-party vendors. These tools typically support the development of analytic functions and models using a variety of techniques and languages such as MapReduce, R and SQL. These functions and models can then be installed and run on the data discovery platform to provide high performance.

Technology Requirements

Given the previous discussion, it can be seen that the technology requirements for a data discovery platform fall into three main categories: data engineering, enhanced analytics and high-performance analytic processing.

In the area of data engineering, the data discovery workbench should provide access to all forms of structured and multi-structured data, and should support flexible data blending, rather than data integration capabilities that require pre-defined and rigid data schemas.

The analytics component of the data discovery workbench should support the many forms of analytic processing discussed in this paper, including OLAP, statistical analysis, text analysis, and predictive modeling and analysis. Some information workers may also require forecasting and optimization tools. The workbench should also provide search and data visualization facilities that make it easy for users to consume and interpret results. The main objective of the analytics component is to provide a flexible set of capabilities that enable rapid and iterative analytics development using a variety of different analysis and visualization techniques.

To aid productivity and ease of use, the workbench should provide pre-built analytic function libraries for processing and analyzing all types of data. For more advanced information workers, systems integrators and independent software vendors, the workbench should include a development kit that enables additional analytic functions to

be added to the system. Examples of analytic languages and programming models that the development kit may need to support include SQL, MapReduce and R.

From a data management perspective, support for high-performance analytic processing is a must. Examples of requirements here include in-database and/or in-memory analytic processing, support for both structured and multi-structured data, optimized storage structures for different types of data and analytic processing (row-based and column-based storage, for example), data compression, hardware exploitation (parallel computing, multi-core processors, processor-specific performance features) and connectivity to external data stores. The data management system should also offer multiple deployment options including software-only, integrated hardware and software appliance, and cloud-based enablement.

GETTING STARTED ON A DATA DISCOVERY PROJECT

The starting point for a data discovery project is not technology selection, but identifying business requirements. Most data discovery requirements are specific to the analytic needs of a specific business area. Data discovery project managers therefore need to work with business unit managers to identify clear business needs and use cases before approaching senior management for funding.

During discussions with business unit managers and senior management, the benefits of a data discovery platform need to be clearly identified. These benefits fall into three areas: the benefits of analytics, the benefits of big data and the benefits of data discovery (see Figure 5).

Technology	Business Benefits
Analytics	<ul style="list-style-type: none">• Reduced costs – stop fraud, reduce risk, identify waste• Increased revenues – improve customer service and satisfaction, increase customer retention and income• Improved efficiency – make better business decisions, react faster to business needs, identify problems before they occur
Big Data	<ul style="list-style-type: none">• Improved access to a richer set of information• Increased insight and knowledge about the business• Faster knowledge creation at a lower cost
Data Discovery	<ul style="list-style-type: none">• Enhanced analytic processes• Optimized business processes and new business opportunities• Flexibility• Faster time to value

Figure 5. Benefits of analytics, big data and data discovery

When getting started on a project and selecting a data discovery platform it is important to:

- Identify the types of data, analytics and workloads to be deployed.
- Determine the total cost of ownership for the system, not just the cost per terabyte of managing the data.
- Understand the skills requirements for deploying and using the platform.
- Investigate if other parts of the organization are using the platform for data discovery.
- Talk to customers who have deployed the platform.
- Understand the impact of the platform on the existing IT infrastructure.
- Be realistic, but pragmatic, about the value and use of big data and data discovery – there is significant business value to be gained from data discovery and big data, but these technologies are still evolving.
- Realize that data discovery and big data are not just about technology, but also about modifying business processes to exploit the value they bring to the business.

VENDOR EXAMPLE: TERADATA ASTER DISCOVERY PLATFORM

Teradata has been developing and marketing database and data warehousing products for more than thirty years. Until the last few years, the company focused primarily on providing hardware and software solutions to organizations that required large-scale and high-performance data warehousing systems. Market dynamics and growth in the use data warehousing and analytics across a broad spectrum of companies of all sizes led the company to expand its Teradata Database portfolio to offer a range of solutions that support a variety of data warehouse sizes and analytics workloads.

The Teradata product set was further enhanced in 2011 with the acquisition of Aster Data Systems. At that time Aster offered a relational DBMS that was optimized for the analytic processing of both structured and multi-structured data. Since the acquisition, the Aster solution has been steadily enhanced to provide a data discovery platform (Teradata Aster Discovery Platform) that can be deployed as a software-only solution, integrated hardware and software appliance (Teradata Aster Big Analytics Appliance), or for use in a cloud environment. For companies that wish to evaluate the Teradata Aster Discovery Platform, a virtual-machine image (Aster Express) is also available for use on a PC.

More recently, Teradata announced a third set of product options (Teradata Portfolio for Hadoop) for managing and processing data in a Hadoop operating environment. Teradata's positioning of its three product sets (Teradata Aster Database, Teradata Discovery Platform and Teradata Portfolio for Hadoop) are discussed later in this paper.

Teradata Aster Product Architecture

The Teradata Aster Discovery Platform (see Figure 6) consists of a data discovery functions library (known as the Aster Discovery Portfolio) and an underlying relational DBMS (Aster Database).

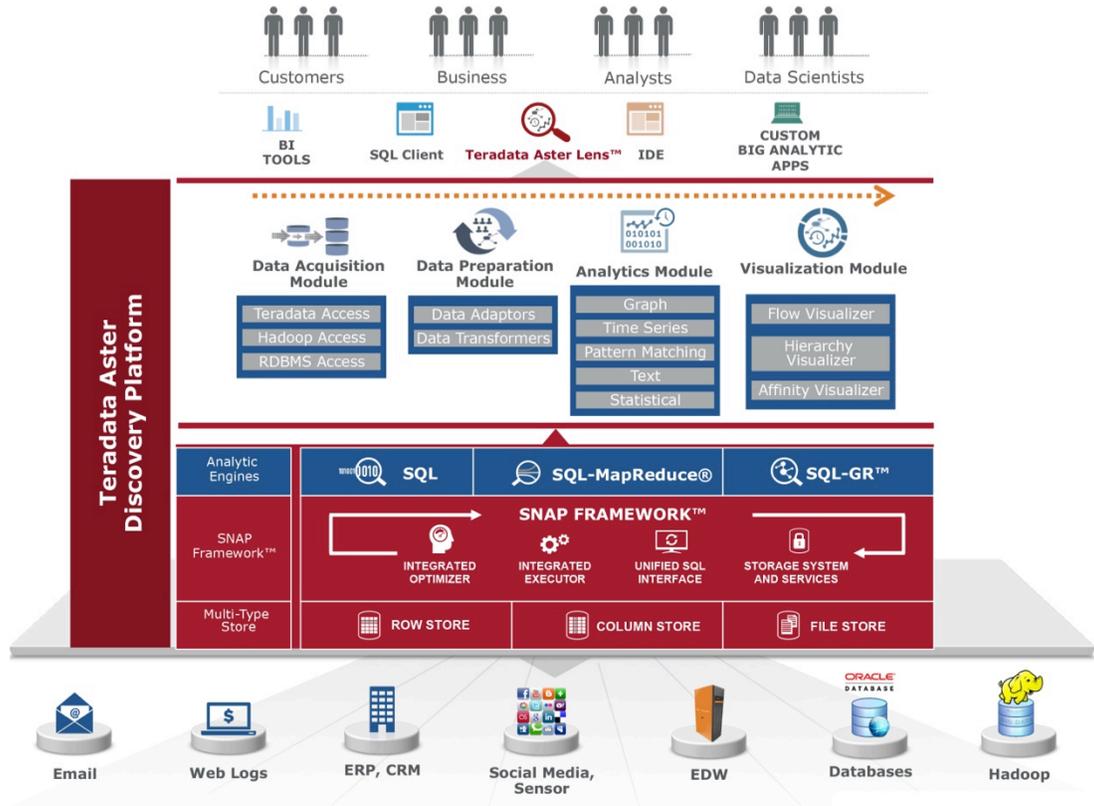


Figure 6. Teradata Aster Discovery Platform

The Aster Database is designed to run as an MPP engine in a shared-nothing storage configuration. The product’s SNAP architecture allows Teradata to add new analytic engines and storage systems that can be integrated easily with Aster’s common and shared services for optimization, query execution, storage management and workload management. The latest release of the product (Aster Discovery Platform 6), for example, adds an analytic engine for SQL graph analysis. Prior releases included analytic engines for standard SQL and MapReduce.

The product supports both row-based and column-based storage of structured data. The latest release adds a file storage engine (Teradata Aster File Store™) that is designed to handle multi-structured data. It is API compatible with Apache Hadoop HDFS. Multi-structured (and structured) data can also be retrieved from external data systems like Apache Hadoop.

The Aster Database functions library contains a wide range of pre-built SQL functions for data acquisition, data preparation, analytic processing and visualization. User-defined functions can be created using a supplied development kit. These SQL functions may be developed using a MapReduce and Bulk Synchronous Programming (BSP) model in a variety of languages, including Java, C, C++ and Python. The open source R statistical language can also be used to invoke MapReduce functions and certain R functions.

Pre-built and user-defined functions can access and process both structured and multi-structured data managed by the Aster Database system. Data discovery applications can also use the Aster Database system to access multi-structured data managed by external data systems (such as a Hadoop system, for example) and transform it (using supplied Aster data preparation functions) into a more suitable format for analysis. The transformed data can be analyzed dynamically *on the fly* by the Aster system, or persisted in an Aster Database for subsequent analysis.

Structured data is analyzed using standard declarative SQL functions, whereas both internal and external multi-structured data is processed by programmatic MapReduce table functions that are invoked using SQL (see Figure 7 for an example).

```
SELECT *
FROM nPath (
  ON (...)
  PARTITION BY sba_id
  ORDER BY datestamp
  MODE (NONOVERLAPPING)
  PATTERN ( '(OTHER_EVENT|FEE_EVENT)+' )
  SYMBOLS (
    event LIKE '%REVERSE FEE%' AS FEE_EVENT,
    event NOT LIKE '%REVERSE FEE%' AS OTHER_EVENT)
  RESULT (...)
) n;
```

Figure 7. Example of the Teradata Aster nPath™ analysis function

Both standard SQL and programmatic functions can be embedded in a single SQL statement – this allows a mixture of structured and multi-structured data to be analyzed by a single SQL query. Both types of functions are executed inside the database system and exploit the parallel processing capabilities of the database engine to enhance performance.

Data Discovery Capabilities

One of the key distinguishing features of the Teradata Aster Discovery Platform is the data discovery functions library and its associated development kit. These data discovery functions are packaged into several licensable *modules*:

Data Acquisition Module: These functions enable access to data managed by a Teradata Database system, Hadoop system or third-party relational DBMS.

Data Preparation Module: These functions process data in web and system logs, XML and JSON files, and Microsoft PST files. Data transforms are also provided for sessionization, pivoting and data de-compression.

Analytics Module: Functions in this module support path, statistical, SQL, text, graph, machine learning and custom-built MapReduce analysis.

Visualization Module: These functions support flow (paths and patterns), hierarchical (relationships) and affinity (clusters and groups) data visualizations. Path analysis could be used, for example, to analyze the events that take place prior to a customer closing an account (see the example in Figure 7). Affinity analysis can help marketers understand what products sell together (market basket analysis). Relationship analysis could be used to analyze abandoned web shopping carts and to drill down into details about specific product categories that lead to abandonment.

Third-Party Modules: Teradata's partnerships include Attensity (for text and sentiment analysis of data stored in an Aster Database), Zementis (for in-database PMML model scoring) and SAS (for in-database analytical processing).

It is important to note that Teradata does not view the Aster Database functions as replacing the need for more advanced data integration and data analysis tools from its partners. The Aster Database functions do, however, provide a valuable environment for data exploration and discovery by less experienced information workers prior to more advanced development of selected high-value candidates by data scientists.

Teradata Strategy

The analytics and big data marketplace is highly volatile at present and how this market will evolve over the next few years is unclear. What is certain is that to exploit the business benefits of advanced analytics and big data, organizations will need to expand their existing data warehousing and decision-making environment to include additional systems that can support usage patterns such as data discovery and data refineries.

Teradata recognizes the trends in the market, and as a result offers three product sets that can be interconnected and exchange data with each other: Teradata Database for enterprise data warehousing, Aster Discovery Platform for exploring and analyzing different varieties of data, and the Teradata Portfolio for Hadoop for managing and transforming large volumes of multi-structured data.

Given the company's history and customer base, Teradata also recognizes that enterprises will be looking to vendors to help make this complex data management environment as easy to access, manage, administer and operate as possible. Teradata's direction here is to provide these capabilities as a part of the Teradata Unified Data Architecture™ (UDA) initiative. The objective of this architecture is to provide an integrated and transparent interface to the Teradata Database, Teradata Aster Discovery Platform and the Hadoop

environment. The architecture helps bridge the gap between the SQL environment preferred by business analysts and the MapReduce and Hadoop environment preferred by programmers. This architecture gives the company the flexibility to evolve with the rapidly changing analytics industry.

I would like to thank Teradata for its support in the publication of this paper.

About BI Research

BI Research is a research and consulting company whose goal is to help organizations understand and exploit new developments in business intelligence, data integration and data management.

EB-6841 > 1013

Aster nPath and Aster File Store are trademarks, and Teradata, Aster and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide.