



CONSULTANCY

Delivering New Business Insights in Big Data Using Teradata Aster Analytics

A Whitepaper

Rick F. van der Lans
Independent Business Intelligence Analyst
R20/Consultancy

January 2016

Sponsored by

TERADATA[®]

Copyright © 2016 R20/Consultancy. All rights reserved. Teradata, the Teradata logo, Aster, SQL-MapReduce, QueryGrid are trademarks of registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Trademarks of companies referenced in this document are the sole property of their respective owners.

Table of Contents

1	Management Summary	1
2	Four Groups Involved in Decision Making	2
3	Increasing the Analytical Strength of an Organization	4
4	Data Scientists and Business Analysts – Two Intertwining Disciplines	6
5	Requirements for an Analytical Solution	8
6	Overview of Teradata Aster Analytics	12
7	How Does Teradata Aster Analytics Support the New Analytical Requirements?	16
	About the Author Rick F. van der Lans	18
	About Teradata Corporation	18
	Appendix A: The Analytical Functions of Teradata Aster Analytics	19

1 Management Summary

The biggest challenge for most organizations is to increase their *analytical strength*. This analytical strength can help them uncover new *business insights* can be uncovered with the goal to survive, to stay cost-effective and competitive, to react quickly, to increase customer care levels, to improve business processes, to personalize products, and so on.

Organizations must increase their analytical strength.

Data scientists, such as actuaries, economists, mathematical technicians, and epidemiologists, are typically responsible for applying sophisticated analytical techniques to uncover new business insights. Commonly, these data scientists have a technical background and have been trained in mathematics and statistics. Unfortunately, data scientists are expensive and there is a considerable shortage of good data scientists. This severely limits the potential analytical strength of an organization.

To make data science affordable for every organization and to deploy data science across the entire organization and not only by the high priests of data science, other disciplines must be involved. It's important that others outside the small group of data scientists are also able to deploy the same set of analytical techniques. The first and most obvious group are the *business analysts*. These analysts must be allowed and must be able to use the same analytical techniques and tools as the data scientists. Undoubtedly, by making the business analysts join the group of people doing data science work the analytical strength of an organization will increase, deepen, and expand.

Business analysts must become involved in data science.

A side-effect is that the boundary between business analysts and data scientists will blur. The two disciplines used to be quite complementary, but in the future the two will be more intertwined. Analytical solutions must support this new group of data scientists. Because business analysts don't have the same background as data scientists, these solutions must be simplified and the analytical techniques must be integrated. This definitely raises the bar for *analytical solutions*. This whitepapers describes requirements for analytical solutions when they are used by data scientists *and* business analysts.

The bar for analytical solutions is raised.

- Business-ready
- Multi-genre advanced analytics™
- Integrated set of analytical techniques
- Data scalability
- Heterogeneous data access
- High-performance analytics

The whitepaper also describes how Teradata's analytical solution called *Teradata Aster Analytics* meets these requirements. It is a SQL-based multi-genre advanced analytical solution for high-end analytics on massive multi-structured data sets. Multi-genre advanced analytics means that a large and integrated set of analytical techniques is

Teradata Aster Analytics is multi-genre advanced analytics at scale.

supported. The product has been designed specifically for analyzing big data to uncover new business insights. Its unique architecture processes analytical techniques inside the database server or directly on Hadoop, leveraging its massively-parallel architecture and analytical engines, such as SQL, SQL-MapReduce, SQL-GR, to fully parallelize the processing of complex analytical queries. In addition, it supports the open-source analytical R engine which comes with a package with over 170 analytic functions and operators to facilitate scalable processing of R scripts. Multi-genre advanced analytics is offered through an easy-to-use SQL or R interface without enforcing analysts to develop in more technical languages, such as Java, Python, and Perl.

Aster Analytics is a key component of the *Teradata Unified Data Architecture* providing multi-genre advanced analytics capabilities. Other components include the Teradata integrated data warehouse, Hadoop, Teradata QueryGrid, and Teradata Listener.

2 Four Groups Involved in Decision Making

The success of organizations depends for a large part on the quality of their decisions. Similarly, the quality of decisions is heavily dependent on the availability of the right data at the right time and in the right form. Within organizations, four groups can be identified that are involved in developing systems that supply decision makers with the right data at the right time: business analysts, data scientists, IT specialists, and business users; see Figure 1.

The success of organizations depends on the quality of their decisions.

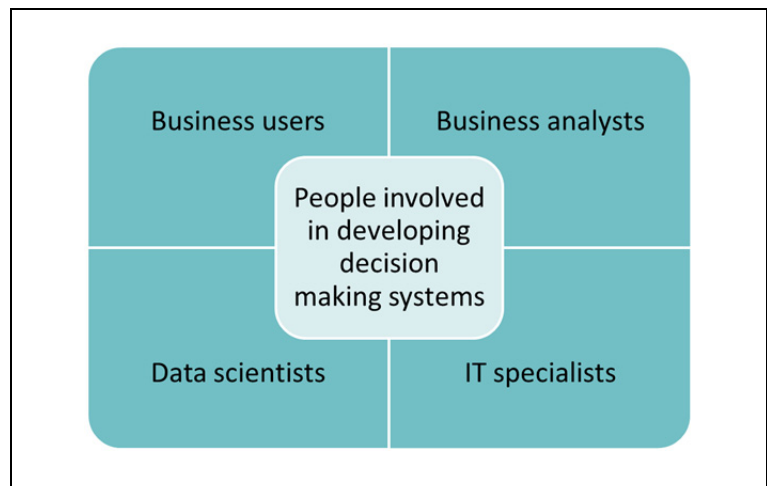


Figure 1 *Four groups are involved in developing systems that supply decision makers with the right data at the right time.*

Business Analysts – Business analysts assist business users in making informed business decisions. They help to transform business needs to reports, to analyze data structures, to define business concepts, and to develop reports for business users. Many of these reports are formally tested, audited, maintained, and governed, and are developed for groups of users. Some reports are developed for regulatory purposes and some for external parties. Quite often, business analysts operate on the border between the IT department and the business departments.

The questions for which business analysts implement reports are often very precise and clear. Examples are:

- Show the revenues per sales region and per product group for the last two years?
- Show how many complaints we receive for each individual brand?
- What is the value of some financial key performance indicator?

The tools used by business analysts are good at reporting and visualization of the data to show business users what has happened within the organization. These tools allow business users to look at data at different aggregation levels, add filters, add dimensions, and so on. With respect to analytical techniques, the tools used are normally restricted to simple forecasting, interpolating and extrapolating techniques.

Data Scientists – Data science is about searching and analyzing data to discover new *business insights* that can lead to business opportunities.

Data science is about searching and analyzing data to discover new business insights.

But what is a data scientist and what does (s)he do? For example, in an oil company the specialists responsible for analyzing soil test results to locate new oil fields or for analyzing new techniques to find new oil fields faster, can be classified as data scientists. An actuary working for an insurance company is another clear example of a data scientist. They deploy mathematics, statistics, and financial theory to analyze the financial consequences of risk. Professors looking for cures for specific diseases by doing DNA research can also be classified as data scientists. Additionally, economists, mathematical technicians, and epidemiologists are also examples of data scientists.

Data scientists have always used a plethora of techniques and tools to do their job. They use different forms of analytics to find the business insights, including forecasting, predictive analytics, and prescriptive analytics. One can say that these techniques together form data science. To quote Wikipedia¹: “Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, chemometrics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing.”

The questions data scientists try to answer are never that straightforward and cannot be answered by simply starting up a particular report or by firing up a pre-defined analysis. The challenge for them is that they don't always know exactly what they are looking for, although they probably have a feeling or inkling. In fact, they have hypotheses that they would like to test out. Examples of questions for which they try to find the answer are:

- What is a possible behavioral pattern of credit card usage that signifies a fraudulent action?
- What are other forms of data that can help us locate deeply buried oil fields more easily?
- How high is the financial risk when a 21 year old person with no job is given a mortgage?

To be successful, data scientists need business knowledge, need to understand the enterprise data, need to know how to deploy the technology and tools properly and efficiently, they must understand statistical

¹ Wikipedia, *Data Science*, October 2015; see https://en.wikipedia.org/wiki/Data_science

techniques, visualization techniques, and most importantly, they need to know how to *interpret* the results. With respect to the tools, data scientists use a wide range of tools from the most advanced ones all the way up to simple spreadsheets, reporting tools, visualization tools, and even occasionally pen and paper.

Recently, the number of new powerful tools for analytics has skyrocketed. Their analytical breath has also improved. What was impossible twenty years ago can now be done by a simple push of the button. Especially with the coming of *big data* so many new technologies have been introduced to research data. For example, using Hadoop technology, data scientists can now analyze data with a speed that was unthinkable a few years ago. This allows them to try out many more alternative hypotheses, ideas, and options, before a new business insight is found.

Many tools for data science don't integrate well.

Unfortunately, this broad set of tools doesn't integrate very well. Results created with one tool cannot be passed easily to the next one. Also, each one has its own unique interface which costs valuable time to learn. And in addition, some of them are difficult to use.

IT Specialists – IT specialists, such as system analysts, programmers, and technical designers, are responsible for developing IT systems that gather data related to the business processes. Examples of such IT systems are inventory management systems, human resource systems, and call center systems. Besides producing new data, these systems also supply numerous business users throughout the organization with data. For example, a call center application allows call center operators to retrieve current contracts of a specific customer, or the inventory management system can show the current stock and dead stock levels of a product to business users. Having this type of data is very valuable for particular business decisions.

Business Users – Business users develop simple reports using simple, easy-to-use, and intuitive technologies such as spreadsheets and self-service BI tools, such as Qlikview, Tableau, and Tibco Spotfire. These tools allow business users to look at data from all angles and at different aggregation levels. For example, a human resource manager uses his own basic overview showing the workshops attended by employees, and a sales person uses his own spreadsheet showing the state of each lead. Some of these tools even support simple forecasting features. Commonly, these reports are simpler than the ones developed by business analysts. Most of the reports developed by business users are for private use only or for a very limited group of colleagues. Note that the simplicity of these reports doesn't imply a restricted value to the decisions they make.

To support their own decisions, business users use simple tools to view data from all angles and at different aggregation levels.

3 Increasing the Analytical Strength of an Organization

For organizations to survive, to stay cost-effective and competitive, to react quickly, to increase customer care levels, to improve business processes, to personalize products, and so on, they must use analytics to uncover new business insights. In fact, they must increase their analytical strength. Therefore, the value of data science to improve the

Organizations must increase their analytical strength.

decisions of an organization is undisputed. The models and information they deliver can make the difference between success and failure.

Unfortunately, of the four groups involved in delivering the right data only two deploy analytics: the business analysts and the data scientists, of which the business analysts are only responsible for a small fraction of all the analytics needed. Most of the real analytics is done by data scientists. This severely limits the analytical strength of an organization, because data scientists are expensive and there is a considerable shortage of good data scientists:

Data scientists are expensive and a scarce resource.

- **Expensive:** Data scientists are usually expensive. Bloomberg Business² reports that the annual starting salaries for data scientists have gone up to over \$200,000. This implies that for many mid-size organizations data science is close to unaffordable, while these organizations can benefit from data science as much as large companies.
- **Scarce:** It can be difficult to find data scientists. Currently, there is a serious shortage of data scientists and this will probably be so in the foreseeable future. McKinsey³ predicts that by 2018 demand for data scientists may be as much as sixty percent greater than the supply. And the Accenture Institute for High Performance⁴ has indicated that there is simply not enough PhD talent to fill the data science jobs. The shortage is especially severe in the U.S. where eighty percent of new data scientist jobs created has not been filled.

To make data science affordable for every organization, and to deploy data science across the entire organization and not only by the high priests of data scientists, a larger group of people must become involved. It's important that people outside the small group of data scientists are also able to deploy the same set of data science techniques. And the first and most obvious group is the business analysts. They must be allowed and must be able to use the same analytical techniques and tools as the data scientists with the intention to improve and broaden the analytical strength of an organization.

Business analysts must become involved in discovering business insights using analytical techniques.

When business analysts become involved in data science, it will also become more affordable for midsize organizations. It won't be a discipline anymore that only large organizations can afford. It will be data science for everyone, which allows every organization to increase their analytical strength.

² Bloomberg Business, *Help Wanted: Black Belts in Data*, June 4, 2015; see <http://www.bloomberg.com/news/articles/2015-06-04/help-wanted-black-belts-in-data>

³ McKinsey Global Institute, *Big Data: The Next Frontier for innovation, Competition, and Productivity*, June 2011; see http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

⁴ Accenture Institute for High Performance, *The Team Solution to the Data Scientist Shortage*, 2013; see https://www.accenture.com/ie-en/~media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Indutries_17/Accenture-Team-Solution-Data-Scientist-Shortage

4 Data Scientists and Business Analysts – Two Intertwining Disciplines

When business analysts become involved in analytics the boundary between their job and that of the data scientists, that used to be clear, will slowly fade. This section describes how the two disciplines, business analyst and data scientist, used to be complementary with respect to the analytical techniques they deployed, and why they will intertwine.

Complementary Disciplines – Once upon a time, data scientists with their sophisticated tools and techniques were seen as magicians who would work their magic in some ivory tower. Occasionally, they would come down to show and explain a new business insight that the organization could use to improve a business process, lower costs, or introduce a more effective way to address customers. Even before the dawn of computers, data scientists were important to organizations to make them, for example, more competitive, to win wars, or to make progress in medicine.

Data scientists didn't work very closely with business analysts. Although both disciplines try to find business insights, for a long time, they used different tool sets, they didn't always use the same data sources, they had different goals, and so on. Due to the need for analytical insights the two disciplines are intertwining.

Data scientists and business analysts may have been using the same data, but they used it differently. The work of data scientists is about searching and analyzing data to produce new business insights that can lead to business opportunities. Business analysts on the other hand develop reports for business users and help them to develop their own reports. These reports can be developed for each layer of the organization, from strategic to operational, and they can be standard reports but also incidental reports. Regardless of this, business analysis and data science must be seen as two complementary disciplines. Together they turn raw data into valuable information and business insights, and are responsible for supplying an organization with the right data at the right time and in the right form. Conclusion, the two disciplines used to be complementary.

Business Analysts Becoming Data Scientists – Normally, business analysts do not use the more advanced techniques deployed by data scientists. Their tools are designed for reporting and visualization of the data to show users what had happened within their organization. These tools allow users and analysts to look at data at different aggregation levels, add filters, add dimensions, and so on. If they need more statistical techniques, they are normally restricted to simple forecasting, interpolating and extrapolating techniques.

This change implies that the tools for analytics must be simplified and the techniques must be integrated. One of the reasons is that data scientists usually get the time to search for new business insights, while for business analysts time can be of essence, because they operate "closer to the business" and must deal with urgent problems for which solutions must be found. Especially analysts operating on the operational level may need answers in minutes.

Evidently, supporting this new group of users has a major impact on analytical solutions. They must make it easy for business analysts to use the more powerful tools easily and correctly. For example, analytical solutions may be using machine-learning techniques to assist the business analyst to pick the right predictive technique for a specific problem, or to propose a visualization form that fits the data to be presented. This helps to make sure the business analysts are deploying the analytical techniques and tools correctly.

Embedding Analytical Models – Another way in which the two disciplines intertwine is when the models created by the data scientists are used by the business analysts. This is done by *embedding* the models in the reports and visualizations developed by the business analysts for the business users. Embedding means that the analytical models are invoked from a standard report. For example, a business analyst needs to develop a report for a group of users that shows trivial data on each customer, such as address, birth date, products purchased, and also includes the risk that this customer will churn within a specific time period. The churn risk is a percentage calculated by a forecasting model. Technically, this means that when the user asks for the data, some of it comes straight from a database, while the churn risk is determined by invoking the analytical model developed by a data scientist.

This embedding of analytical models is very different from “the old days” where one report developed by the business analysts would show the more trivial data and a separate report developed by the data scientists would show the churn risk.

External Data Sources – More and more business analysts and data scientists no longer restrict themselves to *internally produced data* that comes from IT-managed production systems. For their analysis they use all the data they can lay their hands on and that includes external data sources, such as social media data, open data, and public data. Especially with respect to open data, every day more data becomes publicly available. Examples are data sets containing weather-related data, medical data, pollution data, socio-demographic data, crime data, airport data, vehicle collision data, and so on. Especially governments have developed mountains of valuable data. At the time of writing there are 189,920 data sets with US government open data available at the website www.data.gov⁵. Especially tech-savvy data scientists are used to mix this external data with internal data to get the most complete and accurate *business insights*. But it must become easy for the business analysts as well to deploy analytical techniques on internal and external data.

Mixing internal with external data will return the most complete and accurate business insights.

New Data Storage Technology – *Hadoop* is one of the most popular new technologies for storing and analyzing *big data*. Unfortunately, Hadoop has been dominated by data scientist, because they had the technological skills to work with the somewhat technical interfaces of the analytical tools for Hadoop. But as more companies deploy Hadoop in their data architectures, a lot of important data is stored in Hadoop, and now that more business analysts become involved in data science, searching for business insights in these mountains of data must become easy for business analysts without these technical skills.

The Boundary Will Fade – To increase an organization’s analytical strength the boundary between business analysts and data scientists must fade. Figure 2 shows the tools and techniques used by data scientists and business analysts. On the left the fading blue area above the dotted arrow indicates the tools used by business analysts today and below the dotted arrow the ones they use tomorrow.

Data Science for Everyone – To summarize, this realization that the two disciplines are much more intertwined than complementary, may well be the biggest challenge for many organizations in the world of analytics and discovering business insights. Analytical solutions must support this cooperative style of searching for business insights. Also, organizationally data scientists and business analysts must be placed “closer” together. The data scientists must leave their ivory towers and work together with the business

⁵ Data.gov, *The Home of the U.S. Government’s Open Data*, December 2015; see www.data.gov

analysts more frequently and more closely. If done correctly, the entire analytical strength of an organization will increase.

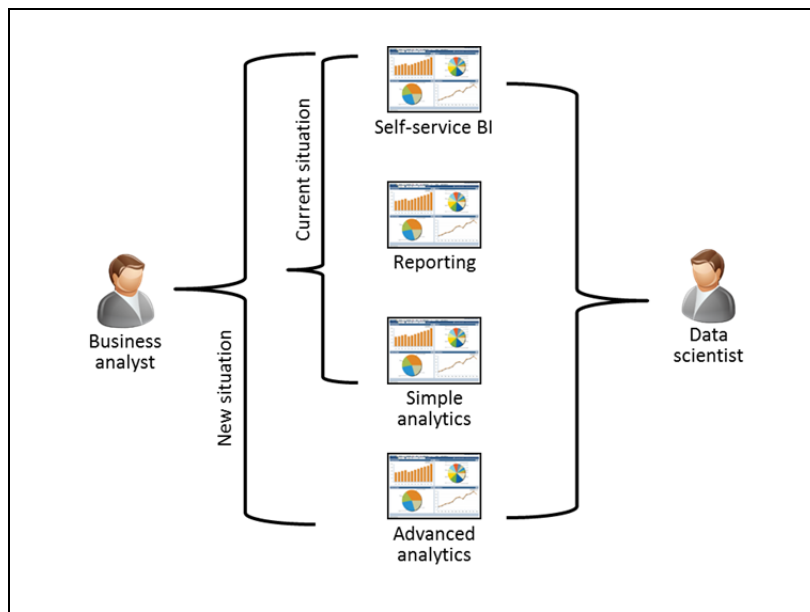


Figure 2 On the left in this figure the current tool usage by business analysts is shown and also the new situation of tool usage. Compare this to the tools used by data scientists as shown on the right.

5 Requirements for an Analytical Solution

By allowing more disciplines to deploy analytical techniques, the bar is raised for *analytical solutions*. This section lists requirements for analytical solutions when they are used by data scientists *and* business analysts.

Business-Ready – Maybe at the time when the analytical tools were primarily used by data scientists there was an excuse for tools to be difficult to use and to be a little Spartan. With more people using the same technology, especially those without a technical or statistical background, the tools *must* be easy to use, well-documented, and supported by the IT organization. There should be no need to be a technical expert to use them. The solution must allow everyone, and especially the business analysts, to apply the techniques commonly deployed by data scientists. In other words, they must be *business-ready*.

Analytical solutions must be business-ready.

Compare this with computers. The first generation of computers consisted of complex machines that could only be managed by specialists wearing cotton gloves and white lab coats. These machines were incredibly complex to use and really needed highly-trained specialists. This changed over time, because more people wanted to use them. The result is that today computers are everywhere and are used by anyone. A key requirement was that the computers had to become much easier to use, like the analytical solutions have to.

Multi-Genre Advanced Analytics – Analytical solutions must support *multi-genre advanced analytics*. This implies that they must support the following two inter-related requirements with respect to analytical

Analytical solutions must support multi-genre advanced analytics.

techniques: first, they must support a wide range of analytical techniques and, second, these techniques must be integrated.

Many analytical solutions support only one or two analytical techniques. For example, they only support graph analytics, text analytics, or offer demographic visualization techniques. In a way, such solutions can be categorized as *one-trick pony solutions*. The analytical demands of today require that data scientists and business analysts can deploy a multitude of analytical techniques, including simple reporting techniques; classic statistical techniques, such as forecasting, predictive modeling, and clustering; data mining techniques; data visualization techniques such as affinity visualization, scatter clouds, geo-visualization techniques; and time-series analysis. There should be no need to switch to other tools when analysts switch techniques, because this would involve too much change. They must learn how to work with all these tools, specifications related to data sets must be replicated for each tool, and so on. Especially for business analysts this is unacceptable.

Multi-genre advanced analytics implies support for a wide range of analytical techniques.

The second requirement is that analytical solutions must allow analysts to *combine* multiple analytical techniques seamlessly for more holistic insights. For example, data scientists must be able to apply graph analytics to analyze a specific data set, then they must be able to use this result coming out of graph analytics in a forecasting technique, and finally they must be able to visualize that result geographically. When doing this, the results must be passed seamlessly from one analytical technique to another. In other words, the analytical solution must not offer a set of standalone techniques, but a set of techniques that can work together: an *integrated set of analytical techniques*.

Multi-genre advanced analytics implies an integrated set of analytical techniques.

For example, when a retailer discovers that their online products are not selling as much as they had forecasted, they want to understand why. Instead of assuming that pricing is the problem they can start looking at their website comments and social media data. They may find many negative feedbacks about the user experience on their site. They can decide to use path analysis to determine how long or hard it is for an online shopper. The analysts of this retailer need multi-genre advanced analytics. Aster Analytics can support this by integrating standard SQL query analytics (sale volumes) with text/sentiment analytics and nPath to study site usability. This is real multi-genre advanced analytics in action.

Also, sophisticated analytical models must be easy to use for business analysts when developing reports for business users. For example, an analyst must be able to use his preferred reporting tool, such as Qlikview or BusinessObjects, to develop a report that shows for each sales region the total sales over the last three years and that also shows for each region the predicted sales using some advanced forecasting technique. It must be a seamless integration.

To summarize, analytical solutions must support multi-genre advanced analytics. They must be like Swiss army knives that can do everything. Multiple analytical techniques must be supported in a coherent, consistent, and integrated way.

Data Scalability – Many traditional information systems store and manage large numbers of records. The last years, big data systems have been developed that store amounts of data magnitudes larger than those in these more traditional systems. For example, click-stream applications, sensor-based applications, and image processing applications all generate massive numbers of records per day. The

amount of records stored surpasses more often than not hundreds of millions of records. According to Google's CEO, Eric Schmidt "Every two days we create as much information as we did from the dawn of civilization up until 2003".

Because big data systems store data on such a detailed level, older analytical tools and techniques used to deploy sampling or aggregation techniques to handle these big data sets. Working with the entire data set would take too long. But sampling and aggregation can hide potential new business insights. Therefore, an analytical solution must allow data scientists and business analysts to analyze all the big data fast and efficiently to uncover business insights that have always been hidden. In other words, an analytical solution must offer *data scalability*. Data scalability⁶ is the ability of a system to store, manipulate, analyze, and process ever increasing amounts of data without reducing overall system availability, performance, or throughput.

Analytical solutions must offer data scalability.

Heterogeneous Data Access – Most users of BI systems find the data they need in the enterprise data warehouse or in one of the data marts. This is not always the case for data scientists, because the data they need can be hidden in numerous data stores of which the data warehouse is probably just one. For example, they may want to include data from external websites in their research, data from one of the many open data sources, results from their own tests and studies, textual data in documents, and so on. Data scientists are *data-greedy*. For them the rule is: more is better, because by being able to analyze more data, more valuable insights can reveal themselves.

The same requirement applies for business analysts when they deploy the same techniques as the data scientists. Therefore, an analytical solution must make it easy to access multiple data stores, including data stores using different technologies. In addition, the platform must allow mixing and matching of data from this heterogeneous set of data stores.

In fact, the platform must present to the business analysts and data scientists this heterogeneous set of data stores as one *logical database*. For them, it's irrelevant which data storage technology is used to store the data. That just slows down analysis. The platform must allow both disciplines to work seamlessly with data wherever it's stored.

Analytical solutions must present all the data as one logical database.

Self-Service BI – With the coming of *self-service BI tools* designed specifically for business users, such as QlikView, Tableau, and Tibco's Spotfire, users were given the *freedom* to develop their own reports. Their intuitive, graphical, and easy-to-use interfaces make it possible for business users with minimal or no IT background to develop their own reports and to analyze the data every way they want. One reason why these tools have become a success is because they help users to analyze data with the speed the business requires.

In the beginning, these products only allowed users to determine and change the visualization of the data interactively. Next, the tools came that gave them more self-service reporting capabilities, and eventually some analytical functionality was added as well.

⁶ Eugene Ciurana, *Getting Started with NoSQL and Data Scalability*, see <http://refcardz.dzone.com/refcardz/getting-started-nosql-and-data>

What is needed is that powerful analytical functionality is added for self-service usage, so that almost every user can use these functions in a simple, intuitive, and interactive way. An analytical solution must provide the users of self-service BI tools the ability to invoke sophisticated analytical functions but with the same ease as with which they invoke the reporting and visualization functions. This would definitely make data science ready for the masses.

Analytical solutions must make data science ready for the masses.

Analytics of Complex Data Structures – Most data values stored in SQL database servers are simple numbers, strings, and dates. But in more and more systems, data has a complex data structure. Examples are weblog entries, sensor messages, text documents, and audio streams. Obviously, these data values do have structure, but that structure is not simple and straightforward. For example, a business analyst may want to analyze an external open data set that contains socio-demographic data. Such a data set may contain its data in a JSON format using all kinds of non-flat structures.

The amount of complex values that organizations store is increasing. If not already, analytics of large sets of complex values will be on everyone's agenda in the near future. It's important that an analytical solution allows data scientists and business analysts to analyze such complex values. Therefore, the solution must make it possible to work with so-called *schema-less* data.

Analytical solutions must be able to work with schema-less data.

High-Performance Analytics – It sounds like kicking in an open door, but for multiple reasons, it's important that an analytical solution is capable of executing its analytical functions fast on big data. In other words, they must support *high-performance analytics*.

Analytical solutions must support high-performance analytics.

The first reason is that data scientists always try to analyze as many alternative models as possible. By adding new variables and changing existing variables, the scientist tries to discover a more fitting and accurate model. When, every time another option is tried, a data scientist has to wait ten minutes before an answer is returned, it stalls the thinking process considerably. They will tend to skip many alternatives purely because of time. They will think twice before they try something out. Therefore, analytical solutions must be able to analyze massive amounts of data fast, because the costs of not finding the perfect model due to poor performance can be high.

The second reason why speed is important is when the models are embedded within the reports of users. Take the customer churn example again. Imagine that every time a user retrieves data on a customer including the churn risk, he has to wait fifteen seconds. For a user with an operational responsibility, this is unacceptable. Imagine this user works in a call center and he has that customer on the phone and he must tell the customer to wait a minute before all the data pops up. Again, this is unacceptable. *High-performance analytics* is not just a nice-to-have feature, it's crucial.

The third reason is related to *real-time applications* where analytics must be executed real-time, no form of delay is acceptable. Take for example streaming applications or some of the systems developed for the *Internet of Things* (IoT). Here, the challenge is that massive amounts of messages are being received and must be analyzed real-time. This definitely requires that the solution supports high-performance analytics.

6 Overview of Teradata Aster Analytics

Teradata Aster Analytics is Teradata's analytical solution for multi-genre advanced analytics on diverse data sets. On the outside, Aster Analytics looks like any other SQL database server. It supports standard SQL and all the common APIs, such as ODBC and JDBC, so that it can be accessed by all the popular analytical and reporting tools. What's inside makes the product special. The product has been designed specifically for analyzing big data to uncover new business insights. Its unique architecture runs analytical techniques, leveraging its massively-parallel architecture to fully parallelize the processing of complex analytical queries. Aster Analytics supports SQL, SQL-MapReduce, SQL-GR, and R to provide analytics with a broad and powerful set of functions.

Aster Analytics is an analytical solution for multi-genre advanced analytics on diverse data sets.

This section describes the key and distinguishing characteristics of Teradata Aster Analytics.

Wide Range of Analytical Functions – Aster Analytics is definitely not a one-trick pony solution, on the contrary, it supports a wide range of analytical functions:

Aster Analytics supports a wide range of analytical functions.

- path and pattern analysis
- statistical analysis (such as histogram, decision tree, correlation, regression, Naïve Bayes classifier)
- graph and relational analysis (such as closeness, pagerank, gtree, and betweenness)
- text analysis and mining (such as sentiment analysis, text categorization, and nGram)
- cluster analysis (such as k-means, canopy, minhash, basket analysis, and collaborative filter)
- machine learning (such as generalized linear model, kmeans, and principal component analysis)

Most of them are available in the form of SQL stored functions that can be invoked from every SQL query. The syntax for these functions is compliant to the ANSI SQL standard. See Appendix A for a complete list of functions supported by Aster Analytics.

The Analytic Engines – Most of the functionality for data science and analytics offered by Aster Analytics is available through the SQL interface for the following analytic engines: SQL, SQL-MapReduce, and SQL-GR.

- **SQL:** As indicated, Aster Analytics supports an ANSI-compliant SQL engine. Because of its support for SQL, which is the most popular database language, any reporting or analytical tool can access Aster Analytics. It's an "open" analytical solution which is not based on some proprietary language. The SQL engine makes Aster Analytics suitable for any type of query workloads, making it an ideal platform for *business analysts*.
- **SQL-MapReduce:** SQL-MapReduce is based on a combination of SQL and a programming model researched by Google called *MapReduce*. The goal of MapReduce is to distribute the processing of analytical functions over as many processors as possible. Analytical functions that are hard to parallelize by most database servers, can be implemented inside MapReduce functions. Values with complex structures and multi-structured records can be processed efficiently. MapReduce functions can be developed that transform the complex values in simple values when the data is retrieved from disk and before it's passed to the applications. Because the execution of these functions is parallelized, schema-on-read is fast.

- **SQL-GR:** The *SQL-GR engine* is a native graph processing engine that allows for powerful graph analysis on large data sets. Possible use cases are social network/influencer analysis, fraud detection, supply chain management, network analysis and threat detection, and money laundering. As the Aster R functions, the Aster SQL-GR functions for graph analytics are implemented as SQL-MapReduce functions and are therefore fully integrated with all the other functions.
- **R Engine:** Aster Analytics also supports the open-source R language, enabling business analysts to run their custom analytic scripts. Because Aster Analytics supports standard R leveraging the Aster MPP architecture, analytical models developed on other R implementations can be ported to run concurrently across multiple compute units within Aster Analytics. Aster's R implementation is not a bare-bone implementation, but comes with a package with over 170 prebuilt analytic functions and operators to facilitate scalable processing of R scripts. These functions convert R commands to SQL or SQL-MapReduce code to run across all the data in the entire database, eliminating the need to partition or sample data.

Aster's R engine is a scalable implementation of the open-source R language.

Integrated Set of Analytical Functions – All the analytical functions are implemented as SQL functions and therefore can be combined. The result of one function can be passed on seamlessly to the next function, as easy as the result of one SQL subquery can be passed on to the next subquery. In other words, all the functions can be mixed and matched. For example, an analyst can first use the function to sessionize weblog entries. Next, the function for sentiment analysis can be deployed on the intermediate result created by the sessionize function, and finally, this intermediate result is visualized through another SQL-MapReduce function. Or, a graph function can be invoked to find the members in a social media network of a specific customer, and next the result is passed to an analytical function to find tweets from those members and apply sentiment analysis on them. Technically, this can all be done with one SQL query.

Note that when combining functions, Aster Analytics doesn't store intermediate results with data after the processing of each function.

Teradata Aster Analytics on Hadoop – A special version of Aster Analytics operates natively on *Hadoop* and is called *Teradata Aster Analytics on Hadoop*. The key difference between Aster Analytics and Aster Analytics on Hadoop is that the former uses its own file system and can access data stored in the *Hadoop Distributed File System* (HDFS). The latter has been designed and optimized to use HDFS as the key file system. The special version runs on Hadoop as a native, first-class engine just like Apache Spark and Apache Hive. Aster Analytics on Hadoop can use the existing Hortonworks and Cloudera Hadoop system.

But the two products only differ in this aspect of data storage. From an analytical viewpoint the two Aster products are identical. Aster Analytics on Hadoop also supports SQL-MapReduce, SQL-GR, Aster R, multi-genre advanced analytics, integrated set of analytical techniques, AppCenter, and so on.

The business-ready nature of Aster Analytics on Hadoop allows a larger community to analyze data stored in Hadoop. Moreover, it supports iterative analytics to rapidly uncover insights. It also allows business analysts and data scientists to use their own preferred analytical tools to analyze Hadoop data. In that way, it protects their existing skillset.

The Teradata Unified Data Architecture – From its first day of inception, Aster Analytics has been an analytical solution for advanced forms of analytics on big data. Since that day the product has evolved and is now a part of the *Teradata Unified Data Architecture (UDA)*. The UDA is a fully integrated set of tools, allowing analysts to use data stored in Hadoop, Teradata, and Aster as if it is all stored in one logical database. Teradata Aster Analytics is a key component of the ecosystem. Other components are the Teradata Database itself, Hadoop, QueryGrid, and Teradata Listener. A high-level overview of the Teradata UDA is presented in Figure 3.

Teradata Aster Analytics is an integrated part of the Teradata Unified Data Architecture.

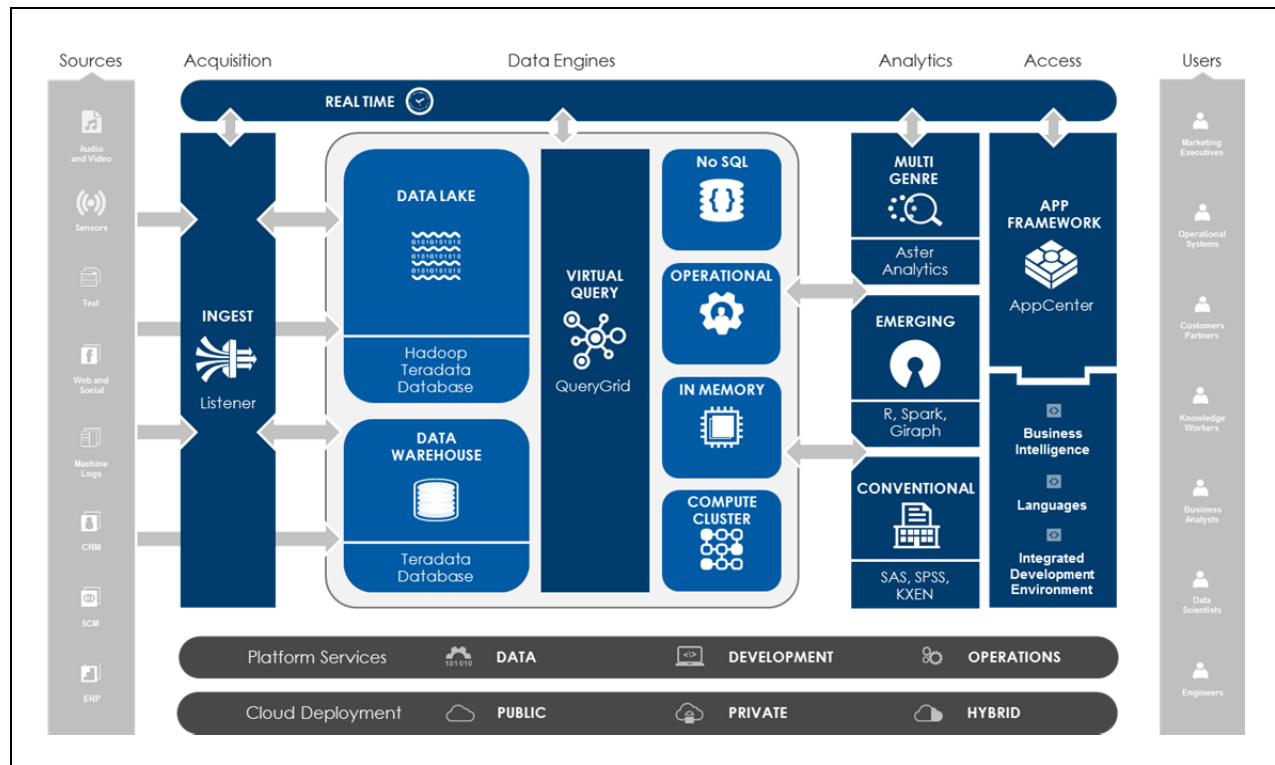


Figure 3 *The Teradata Unified Data Architecture includes the Teradata Database, Teradata Aster Analytics, Hadoop, QueryGrid, and Listener.*

Teradata QueryGrid is the traffic cop in this eco-system. If Teradata or Aster Analytics needs data from one of the other data sources, it asks QueryGrid to get that data. QueryGrid is also responsible for federating the data from multiple databases, so, if an application wants to join a table in Hadoop with a table in Aster, QueryGrid makes sure the join is processed efficiently.

Teradata QueryGrid is the traffic cop of the Teradata Unified Data Architecture.

For the applications and users QueryGrid hides the distribution of tables across the three platforms. For example, a query can be defined in Aster Analytics that points to a table stored in Teradata or in Hadoop. For users there is no difference when this table is queried or a table whose data is stored natively in Aster Analytics. Aster Analytics knows that the table being accessed is stored somewhere else and passes the query to QueryGrid which takes care that the right data source is accessed.

In this process, Aster Analytics and QueryGrid together try to “pushdown” processing when it makes sense. They try to execute as much of the query operations on the data source itself with the intention to minimize the amount of data movement between systems. This can be done for many operations, but not for all. Especially, many of the advanced analytical functions supported by Aster Analytics cannot be executed by Teradata or Hadoop. In this case, the data needed for the functions is extracted from Teradata or Hadoop and copied to Aster Analytics and then the execution of the functions is done by Aster Analytics itself.

Teradata Database users are also allowed to invoke the advanced analytical functions of Teradata Aster Analytics. The Teradata Database recognizes that Aster Analytics functions are invoked and it requests Teradata Aster Analytics to execute these functions. This may mean that data must be moved from Teradata or Hadoop to Teradata Aster Analytics as well. To the users it feels as if all the processing was done by Teradata itself and on data stored in a Teradata database. All this processing is seamlessly done across the eco-system in which QueryGrid plays a key role.

The Teradata Aster AppCenter – The SQL interface, the extensive set of analytical functions, and the ability to combine functions, allow for almost unlimited forms of analytics. With *Teradata Aster AppCenter* all this functionality can be hidden in pre-built analytical applications; see Figure 4. Data scientists and business analysts can develop a sophisticated analytical model consisting of complex formulas, functions, SQL statements, and Java code. With AppCenter all of this logic can be wrapped into a simple application. Such apps can be made available to all kinds of users to invoke their analytics. Via the web-based interface users can invoke the apps via a web browser on many devices, including tablets and smartphones. AppCenter makes invoking analytical functions very easy. In fact, the wrapped functionality can be invoked by everyone. AppCenter supports self-service analytics.

With AppCenter all the analytical power of Teradata Aster Analytics can be made available to business users through simple-to-use apps.

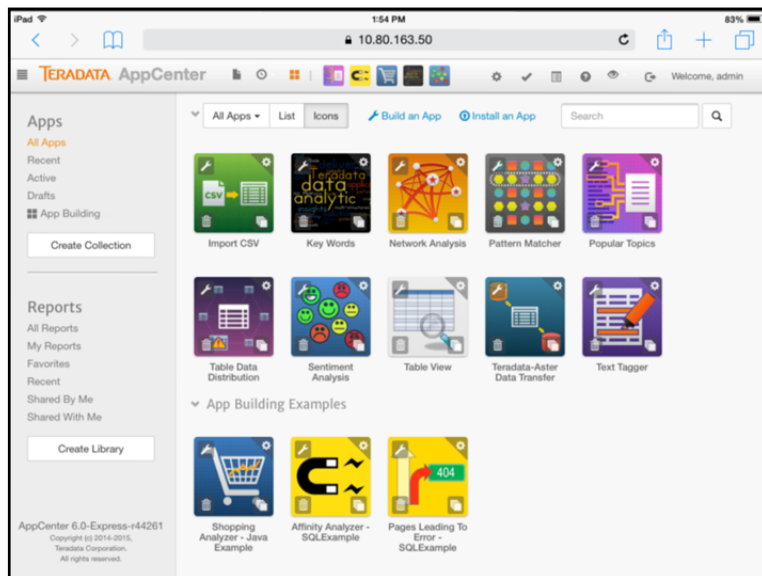


Figure 4 *With AppCenter complex analytical functions can be made available to business users through a simple interface.*

Aster Analytics also includes the ability to package models and all associated data transformations into implementation modules that can be applied in real time. For example, a customer churn model that creates a likelihood score for each customer can be applied as new raw data streams in without the need

for any separate data transformations that would feed data into the model. All such transformations are automatically included with the actual model and applied in real time to create a likelihood score, based on which further decisions on the customer can be taken.

7 How Does Aster Analytics Support the New Analytical Requirements?

This section describes how Aster Analytics supports the requirements for a modern analytical solution as listed and explained in Section 5.

Business-Ready – There are multiple ways in which Aster Analytics is business-ready and qualifies as a platform for data scientists as well as business analysts. First, with AppCenter the most complex forms of analytics can be wrapped in easy-to-use apps that can be invoked by business users and business analysts on their preferred devices. Second, because analytical functions are executed from within SQL statements, they can be invoked from any reporting or analytical tool capable of invoking SQL statements. For example, a simple report developed by a business analyst in Microsoft Excel can invoke the most complex analytical functions supported by Aster Analytics. This is all handled transparently to the business users.

Especially Aster AppCenter makes Aster Analytics business-ready.

Multi-Genre Advanced Analytics – Aster Analytics' implementation of SQL, SQL-MapReduce, Aster SQL-GR, and R offers almost every genre of analytical function imaginable; see Section 6. Teradata Aster Analytics is not a one-trick pony analytical solution; it really is a multi-genre advanced analytical solution.

Aster Analytics supports multi-genre advanced analytics.

If needed, business analysts and data scientists can extend the list of analytical functions by developing their own. And if they are coded correctly, they become *polymorphic* allowing code to be written independent of the tables and columns being accessed. The advantage of polymorphic functions is that the same type of function doesn't have to be written separately for every table and column. For example, a function can be written that determines the top ten values of a column and it can be invoked for every column of every table. This polymorphism allows specialists to develop a large set of specific functions that can be invoked by every business user, report, BI tool, and from every app developed with AppCenter.

Integrated Set of Analytical Techniques – As indicated in Section 6, Aster Analytics offers the ability to mix and match all the analytical functions. The results of one analytical function can be passed on to the next. With some of the other tools, when analysts first want to use graph analytics followed by text analytics, they are forced to take the results of the first analysis and copy that to another text tool to deploy text analytics on the graph result. Especially when big data is involved, this somewhat clumsy export and import of results can be a tedious and error-prone exercise.

Data Scalability – The data sources of the Teradata Unified Data Architecture, Aster Analytics, Teradata Database, and Hadoop, are all designed to offer data scalability. They were all designed for big data even before the term existed. All of them have proven their value in massive data environments. This means that all the analytical functions supported by Aster Analytics can be executed on these scalable platforms.

Aster Analytics was designed for big data before the term even existed.

Heterogeneous Data Access – QueryGrid makes heterogeneous data access possible. It hides the fact that the data to be analyzed is stored with different data storage technologies (Teradata, Aster Analytics, or Hadoop). For every user of Aster Analytics and Teradata, it's all one big logical data source. When data from multiple data sources has to be joined together, QueryGrid federates them seamlessly. In other words, QueryGrid turns all the big data stored in the various data stores into one scalable big data environment.

Self-Service Analytics – AppCenter offers self-service analytics. As indicated in the previous section, with AppCenter all Aster's functionality can be hidden in pre-built analytical applications. With AppCenter all the complex formulas, functions, statements, and Java code, can be wrapped into a simple application. Such apps can then be made available to all business users to invoke analytical functions.

Analytics of Complex Data Structures – Although Aster Analytics' analytical functions are invoked via SQL, it efficiently processes values with complex data structures. MapReduce functions can be developed that transform the complex values to simple values when the data is extracted from disk and before it's passed to the applications. This flattening of data is fast, because the execution of these functions is parallelized. Conceptually, the way this works in Aster Analytics is very much like how MapReduce and HDFS process complex values. In addition, Aster Analytics supports schema-on-read functionality. Data with no clear structure and schema can be turned into data with a schema before the other functions are applied to the data.

Teradata Aster Analytics processes complex data structures and schema-less data efficiently.

High-Performance Analytics – Aster Analytics' internal architecture has always been a perfect match for MPP hardware architectures. For example, the processing of SQL operations, such as joins, group-by's, complex calculations, and operations that are non-relational by nature including most of the time-series based operations is always parallelized. Moreover, due to predictable query performance, the environment scales almost linearly.

About the Author Rick F. van der Lans

Rick F. van der Lans is an independent analyst, consultant, author, and lecturer specializing in data warehousing, business intelligence, big data, database technology, and data virtualization. He works for R20/Consultancy (www.r20.nl), a consultancy company he founded in 1987.

Rick is chairman of the annual European Enterprise Data and Business Intelligence Conference (organized annually in London). He writes for Techtarget.com⁷, B-eye-Network.com⁸ and other websites. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles⁹ all published at B-eye-Network.com. The Data Delivery Platform is an architecture based on data virtualization.

He has written several books on database technology. Published in 1987, his popular *Introduction to SQL*¹⁰ was the first English book on the market devoted entirely to SQL. After more than twenty-five years, this book is still being sold, and has been translated in several languages, including Chinese, German, and Italian. His latest book¹¹ *Data Virtualization for Business Intelligence Systems* was published in 2012.

For more information please visit www.r20.nl, or email to rick@r20.nl. You can also get in touch with him via LinkedIn and via Twitter @Rick_vanderlans.

About Teradata Corporation

Teradata (NYSE: TDC) helps companies get more value from data than any other company. Teradata's leading portfolio of big data analytic solutions, integrated marketing applications, and services can help organizations gain a sustainable competitive advantage with data. Visit teradata.com for details.

⁷ See <http://www.techtarget.com/contributor/Rick-Van-Der-Lans>

⁸ See <http://www.b-eye-network.com/channels/5087/articles/>

⁹ See <http://www.b-eye-network.com/channels/5087/view/12495>

¹⁰ R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.

¹¹ R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.

Appendix A The Analytical Functions of Teradata Aster Analytics

This appendix contains a list of functions supported by Teradata Aster Analytics. It shows the richness and the extensiveness.

<p>Statistical Analysis Approximate Distinct Count Approximate Percentile Average Confusion Matrix Correlation CoxPH, CoxPredict, CoxSurvFit Distribution Matching Exponential Moving Average Enhanced Histogram F-Measure Generalized Linear Model (GLM) GLM Predict Histogram K-Nearest Neighbor Classification (KNN) LARS Functions Linear Regression Logistic Predict (deprecated) Logistic Regression (deprecated) LRTEST Percentile Principal Component Analysis Sample Simple Moving Average Support Vector Machines Vector Distance Volume Weighted Average Price Weighted Moving Average</p> <p>Naïve Bayes Naïve Bayes Map, Reduce & Predict</p>	<p>Data Transformation Antiselect Apache Log Parser FellegiSunter Trainer & Predict E-Mail Parser Identity Match IpGeo JSON Parser Multicase Murmurhash Outlier Filter PST Parser AFS Pack Pivot Scale Functions String Similarity Sessionization Unpack Unpivot URI Pack & Unpack XML Parser & Relation</p> <p>Cluster Analysis Canopy Gaussian Mixture Model KMeans KMeansPlot Minhash</p> <p>Association Analysis Basket Generator Collaboration Filter WSRecommender</p> <p>Decision Trees Forest Drive, Predict & Analyze Single Tree Drive & Predict</p>	<p>Graph Analysis All Pairs Shortest-Path Betweenness Closeness Eigenvector Centrality gTree Hidden Markov Model Local Clustering Coefficient Loopy Belief Propagation Modularity nTree PageRank Personalized SALSA Random Walk Sample Shapley Value</p> <p>Text Analysis Chinese Text Segmentation Latent Dirichlet Allocation (LDA) Levenshtein Distance Naïve Bayes Text Classifier Named Entity Recognition (NER) nGram POS Tagger Sentenizer Sentiment Extraction Term Frequency-Inverse Document Frequency (TF-IDF) Text Classifier Text Chunker Text Morph Text Parser Text Tagging Text Tokenizer</p>	<p>Time Series, Path & Pattern Matching Arima & Predictor Attribution Burst Cumulative Moving Average Dynamic Time Warping (DTW) Discrete Wavelet Transforms Discrete Wavelet Transforms on 2 Dimensional Input Frequent Paths Inverse Discrete Wavelet Transforms Inverse Wavelet Transforms on Multiple Sequences Interpolator nPath Path Generator, Starter & Summarizer Symbolic Aggregation Approximation (SAX) Sessionization Shapelets</p> <p>Location Analysis Load Geometries Point in Polygon Geometry Overlay</p> <p>Visualization CfilterViz NpathViz</p> <p>Aster Database System Function nc_skew nc_relationstats</p>
--	--	--	--