A BTI Study

Evolution of Big Data Analytics: Experiences with Teradata Aster and Apache Hadoop

Richard Hackathorn, Bolder Technology Inc.
March 2013

This study explores the evolution of big data analytics and its maturity within the enterprise. The initial focus was on the approaches and economics to using Teradata® Aster Discovery Platform and Apache Hadoop within the same analytical architecture. Three companies anonymously shared their experiences and practices with big data analytics and highlighted the benefits and issues of their dual Aster-Hadoop architectures.

The study also outlines the tradeoffs and insights of how the practice of big data analytics is maturing in companies today. Companies are making substantial investments in analytical architectures and must be knowledgeable of these tradeoffs. The technology is moving very fast and, hence, companies must satisfy current requirements while constantly planning for future requirements. The take-away of this study is a series of insights that a company should periodically debate as its analytical architecture matures.

Challenges

This study investigates the challenges facing professionals who are blending big data and analytics within their enterprises. In particular, the focus is how to determine the best analytical architecture that will serve both the current and future demands of their company. Amid the rapid changes in technology and its applications, several critical tradeoffs should drive the evolution of this architecture.

Big data is big! Analytics using big data is hot! The new technologies for big data analytics are incredible and evolving rapidly. New business applications are emerging weekly and are revolutionizing the way that many industries do business with their customers, suppliers and other partners. This summarizes the feelings voiced in many industry publications. However, what is really happening? What are the cutting-edge companies doing today?

In this study, the term 'big data analytics' is defined as:

> Analytics[1] using big data (as characterized by volume, velocity, and variety[2]) within an enterprise architecture (across multiple functional areas) to support critical operational processes (as contrasted with one-time ad-hoc analyses).

In the past, analytics was reserved for back-room deliberations by data geeks generating monthly reports on how things are going. Today, analytics make a difference in how the company does business, day by day, and even minute by minute. The analytical applications are 'mission-critical' components in the overall business processes. If the analytics break, executives and customers are upset! This is a dramatic change in the role of big data analytics.

In this study, the term 'analytical architecture' refers to the information technology (IT) architecture used to support big data analytics in a company. A similar term is 'discovery platform' to emphasize the data discovery process (of predictive analytics, data mining and the like), as contrasted with the pre-defined reporting and analysis processes in BI suites. Further, the term is used in this study to distinguish between the larger and more inclusive 'enterprise architecture.' There are several issues discussed later about the relationship between the two types of IT architectures.

To understand these changes, we interviewed three innovative companies, in different industries, who are pioneers with big data analytics. From these interviews, each of the companies mentioned the following challenges.

Tale of Dual Platforms

The first challenge is to compare and contrast two platforms suitable for processing big data volumes, which are Apache Hadoop[3] and Teradata Aster[4]. There are many companies that have one but not both. To solicit knowledgeable opinions, the study selected companies that have incorporated both platforms into their analytical architecture. The objective is to document how the companies leverage the unique strengths of the platforms to generate business value.

It was apparent that each company appreciated both platforms for a variety of reasons and found complementary in several ways. In general, the opinion was that Hadoop provided a creative and increasingly stable mechanism to acquire and retain big data, while the Teradata Aster platform provided better productivity for discovering patterns and exploring data. Both overlap in functionality when filtering and transforming data. The release of the Teradata Aster SQL-H™ connector was pivotal in shifting toward an architecture that blends both platforms. More details are given in the three case studies.

Data Tsunami

The second challenge is to turn the fears of the incoming data tsunami into realistic expectations about the benefits of big data analytics to the company. For instance, the interviewees often discussed the volumes and velocities of acquired data. In some cases, companies are acquiring several terabytes daily, implying that the total data store must handle petabytes to maintain a multi-year history. In addition, there were interesting discussions centered on new data sources to support new business applications. Further, the (often devilishly) multi-structured nature of the data is driving analytics to higher levels of sophistication.

From the interviews, it was apparent that the tsunami analogy is becoming inappropriate. Instead of a large remote force flooding a company with overwhelming data, companies are finding big data opportunities in every corner of their business. Lift a rug in any room, and there are terabytes awaiting analysis!

Data to Action

The next challenge is to document the ways that companies managed the analytical processes throughout the entire analytics value chain from raw data to business actions, as illustrated in Figure 1.

Figure 1 - The Analytic Value Chain.[5]

Analytics generate business value only when it can improve business processes through specific changed actions. In the past, the emphasis was on generating information, which was distributed to the 'right' people in a timely fashion. The assumption was that those people would consume this information and perform their job functions in a more effective (or at least efficient) manner. However, things happen too quickly in today's companies for this paradigm to be sufficient. Analytics must be integrated directly into business processes, with the proper human oversight. With big data, the analytical architecture must focus on how structure (meta-data) enhances the raw data, illuminating the myriad of relationships that link one data element to another. That is why cross-functional data linkages into the corporate data warehouse are critical for a viable analytic value chain.

Total Cost of Ownership

The final challenge concerns the total cost of ownership (TCO) of an analytical architecture. The challenge is determining the proper method for calculating the total cost to acquire, maintain and extend the analytical architecture for the company. In the interviews, it was difficult to obtain specific amounts for the various costs associated with analytics. However, there was considerable opinion on what cost factors a company should consider, such as use of current technical skills, flexibility in prototyping new applications, pay-as-used cloud-based services, and depth of analytics. In other words, a serious TCO investigation for an analytical architecture will involve cost factors more diverse that a traditional IT architecture. Companies that mandate an adherence to a traditional IT assessment will be at a disadvantage to pursue innovative applications involving big data analytics. The responsible executives must think strategically and broadly about TCO assessments of potential analytical architectures.

- - - - -

To understand these challenges, the next section describes how these companies applied big data analytics in their businesses:

- Company A – Innovative e-Commerce Retailer
- Company B – Global Advertising Agency
- Company C – Large Healthcare Provider

Company A – Innovative e-Commerce Retailer

Company A is an innovative e-commerce retailer that sells a diverse set of products and services. Their distinctive is to provide their customers with an exciting online shopping experience tailored to their needs, along with superb service. The success factor for Company A is to know more about their customers so they can cater to them better.

The Director of Data Engineering for Company A was interviewed. He has a team of twelve persons, whom are part of a technology group supporting data warehousing, business intelligence, and analytics. He described their focus as, "*ensuring that everyone in the company has better access to the relevant data for marketing optimization and the A/B testing platform*." His group is… "*Growing very well and am totally psyched about it.*"

Background

He continued by explaining the business problem and their approach…

"Our goal is to be more responsive to the customer – to have a finger on the pulse of what's going on. By integrating and analyzing data about the website behavior of customers, we can draw relationships that others may not have seen. It is more than data integration; it is pattern recognition."

His team integrates clickstream information with email logs, ad viewing, and operational data "to figure out what is going on with our customer." This data was used to support A/B testing and multivariate testing to track the entire customer's experience – from searching for the

product, ordering it, receiving the package, and even product returns, all to optimize the experience for the customer.

Architecture

He described the evolution of their architecture over the past two years…

"The company was just a hundred people two years ago. We had no analytic tools. Our team was just two people and me. Given our tradition of doing things in an innovative way, we started using MapReduce with Apache Hadoop for analytics. However, we knew that this architecture would not be sufficient. We needed to build a data warehouse. In addition, Cognos and Spotfire were previously purchased and had a user base that required support. So, we investigated other analytic platforms, such as Aster Data (before its acquisition by Teradata), Oracle, Vertica, Greenplum, and MongoDB."

About a year ago, the team acquired Teradata Aster to complement their Hadoop platform. The Teradata Aster platform is being used as a reporting platform, along with its analytic processing for discovering new insights beyond reporting.

When asked about the reasons for deciding on Teradata Aster, he explained, "*I explain what I need to the Aster folks. We immediately had a good working relationship with some quality Aster folks. The project quickly gelled into place and has worked successfully over the past year*."

Teradata Aster Database allows the team to analyze more of these potential behavior patterns in a stable and scalable way. In particular, the distributed parallel nature of the Teradata Aster discovery platform allows reasonable computation times with large data sets.

In Figure 2, the architecture shows the data sources, analytic platforms, and data delivery, from left to right.

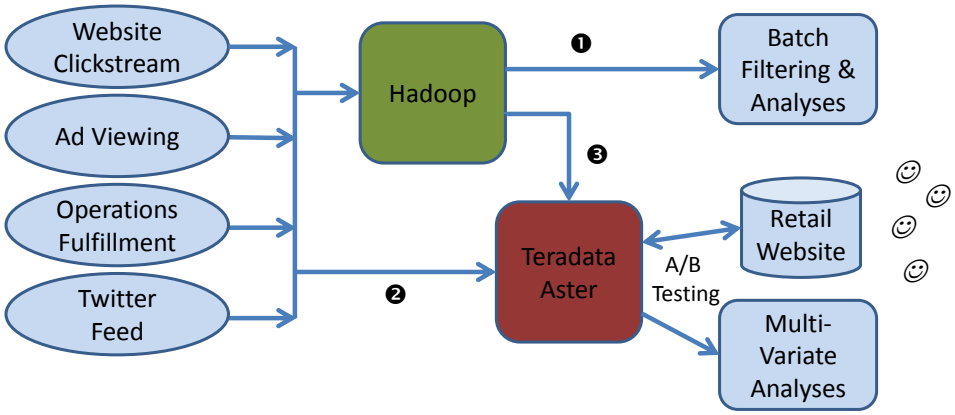


Figure 2 - Architecture for Company A.

As indicated, there are three paths through the architecture. First, there is a direct load into the Hadoop platform for ad-hoc analyses using Hadoop MapReduce. Second, there is filtering and cleansing processing into the Aster platform, which drives the A/B testing for the retail website.

Third, there is a direct load into the Hadoop platform as cold storage, and then data is extracted and loaded using Hadoop MapReduce into the Aster platform for the analytical processing.

The Hadoop cold storage contains data that is infrequently accessed. By using Amazon EC2, the company is able "to absorb large data volumes with fluctuating demands" and then, as needed, transfers data sets to Amazon Elastic MapReduce. He offered some general observations about the Hadoop platform…

"Hadoop Distributed File System is the sexy part of the Apache Hadoop project. It is open-source and allows the problem solving qualities of MapReduce to really shine. It is a brilliantly simple framework that works well."

On the other hand, he concludes that the Aster platform is good at discovering new insights in the data since, as he remarked, "*Analytics, like R, do not behave nicely in Hadoop.*"

Raw data is extract from various sources and loaded onto Hadoop for initial exploration. Later, the data is filtered/transformed and retained for 6 months on the Aster platform for rapid ad-hoc analysis. He explained that not all of the raw data is stored on the Aster platform, mainly for cost reasons. *"Not every piece of data flows to Aster, since we need to do some digging ahead of time using MapReduce in Hadoop.*"

His team has developed many MapReduce functions in the Aster SQL-MapReduce® framework, which is much easier than in standard SQL. One analytic function that they support on the Aster platform is a random forest analysis[6] used in several business areas.

He described the Teradata Aster platform is "our Swiss army knife" within their IT architecture, since it enables his team "*to cover all the bases needed to solve the business problem*." He noted that the Aster platform has the flexibility to manage a variety of data types. In addition, it has linkages to the Hadoop platform and support for the newer analytic tools like Spotfire.

Although successful in some areas of analytics, he feels that the industry is now finding that Hadoop does not cover a wide enough breadth of analytical processing for an enterprise. "*The Hadoop platform cannot do all the things I need. However, I have the Aster platform to cover those areas*." Over time, he will shift more of the workload onto the Aster platform, using the Hadoop platform less for analytic work and more for mass storage.

He then offered three examples of recent projects that utilized these features:

- *"We did an analysis of browser usage by Internet Explorer IE6 against our website. The raw clickstream was initially stored in Hadoop, summarized, and then extracted into the Aster platform for use with Spotfire and R. A valuable feature of the Aster platform is the compatibility to output MapReduce results to Cognos with ODBC connectors, Spotfire with JDBC connectors, or standard SQL queries.*
- *"A really cool application was a MapReduce program that hits Google analytics as a web service, pulling its information into the Aster platform. Then, we used SQL-MapReduce to query Google analytics with the proper joins, without permanently storing this data. In addition, the application executes within a reasonable timeframe. The developer wrote this application in a few days. We cannot do this with Oracle or other traditional tools.*
- *"Another application that I like was… We tokenize and parse large volume of text from Twitter feeds to gage sentiment toward the company.*

Cost Factors

His team had considerable experience with the PostgreSQL database (from which the Aster platform was derived). Hence, the adoption of Teradata Aster was an easy transition for our developers and database administrators, thus avoiding the cost of acquiring specialized database skills.

Because of cost factors, the Aster platform only retains data used by the analytics, which are a few terabytes rather than petabytes. "*We are a large company so the IT costs are critical*." In contrast, he was more concerned with the support costs of Cognos, which "*demands a lot of infrastructure.*"

The adoption of Amazon EC2 is easy for its incremental cost structure and reasonable technical skills. Provisioning Hadoop via the Amazon Elastic MapReduce is also an easy migration path for his team. Hadoop cluster needs at least six machines but they can ask for 40 machines if the MapReduce job requires it. However, Hadoop has long latency times to 'spin up.' He remarked, "*Hadoop takes time to start and do its work."*

He noted that Hadoop MapReduce can "*solve many problems, but you need a Ph.D. to parallelize the processing efficiently,*" implying that analytics with Hadoop require expensive and scarce persons with a deep skill set. Further, it is hard to find persons with the skills to maintain the Hadoop platform and to avoid the costs of enterprise support for Hadoop from Cloudera or Hortonworks.

Lessons Learned

He was asked to share some lessons that his team has learned over the last two years. He mentioned the following points:

- Design a proof-of-concept (POC) exercise using a specific problem set to be solved by Hadoop, Aster, or another analytic platform. Prove to yourself whether each platform can provide the proper answers for that problem set and can handle its estimated workload.
- We decided not to create our own data center with on-site hardware. Instead, we are using Amazon EC2 services where we can provision extra resources only when needed. Thus, we have avoided the usual problems and costs associated with operating a data center. And, we avoid paying on-going operational/support costs that are configured for periods of peak demands.
- The Aster platform behaves like a typical database system. We are not worried, or limited by its infrastructure. Database operation can be assigned to a normal database administrator. In other words, "*we get it!*" Moreover, we are not worried about the entire system going down.
- Any new and innovative application of new technology will have its problems. Expect it. Be patient. Your vendor should collaborate with your company to resolve those problems, as Teradata Aster did with us.

Summary

In just two years, Company A matured their IT architecture in data warehousing and in analytics in innovative ways. Cost factors like time-to-value and risk assessment weighted heavier than staffing rates and software licensing. The use of the Teradata Aster platform was the 'Swiss army knife' in his toolkit, allowing him to cover many data analysis requirements with minimal infrastructure. And, the use of Hadoop in Amazon Web Services (AWS) provides continuing business value as 'cold storage' for the big data tsunami.

Company B – Global Advertising Agency

Company B is a large full-service advertising agency with digital marketing and technology at its core. The company creates digital media that build business identity through web development, media planning and buying, technology and innovation, emerging media, analytics, mobile, advertising creative, social influence marketing, and search. Their clients are some of the largest global corporations.

Background

The Vice President of Business Intelligence was interviewed. He described his company and its use of big data analytics as follows:

"Our core usage today is focused on a multi-attribution model that emphasizes the segmentation of media audiences. In contrast to the 'last touch' approach, we are now analyzing a comprehensive view of all the touch points of customers across all multi-media channels, such as interest on an advertiser's website in addition to behavior on the vendor's on-site website. For instance, when customers received an email, do they look at it, and what do they do, such as click for additional information? And, when customers search on related topics, do they click through on displayed ad banners? And, when a customer purchases a product, what was the sequence of touches that led to that purchase?"

The heritage of Company B is ad serving. In other words, their focus is optimizing multi-channel digital marketing by understanding customers' behavior across all media channels, enhancing this data with third-party information, and assessing impact that specific media has on driving value to the business. In addition, they segment the customers based on how media influences their actual purchases, thus optimizing the targeting of ad channels.

This customer-segmentation analytics provides critical information to the company's internal teams who support clients. They manage the whole range of services to their clients, from analyzing the on-line experiences of their customers, purchasing the media, and managing Search Engine Optimization (SEO) work. "*The company covers the whole gambit!*" The internal client teams use the insights emerging from the media segmentation analytics to optimize digital marketing campaigns from the client. The bottom line for value to our clients is to optimize how they spend their media funds on digital marketing.

Architecture

Company B processes 5.5 terabytes or 36 billion rows of clickstream data every day from digital media, site behavior, social media, and offline media. In media speak, that's about 400 billion media impressions per year.

The daily process uses Amazon Elastic MapReduce (EMR) to cleanse and aggregate the clickstream data into transactional cookie-level (or session-level) data, which passes to Aster Database for advanced analysis. In addition, the large data sets are retained in Amazon S3 cold storage for future analysis.

The transactional data is partitioned among their clients. Clients have their separate custom data marts so that the client data is not co-mingled. The client data is optimized and structured as custom OLAP cubes for use by our internal client teams. Reporting to the clients about digital marketing performance is performed from these cubes.

Company B built a data mart customized to the unique business requirements for each client. Access to the client data mart is primarily by our internal teams. Though there is high-level reporting directly to clients so that they can understand why we decided on certain advertising optimizations. Clients need to know why Company B has moved their funds from one advertising channel to another.

In addition, the Aster Database platform performs advanced analyses on the total transactional data set to generate the customer segmentation definitions, which is used for media targeting at the customer instance by each client.

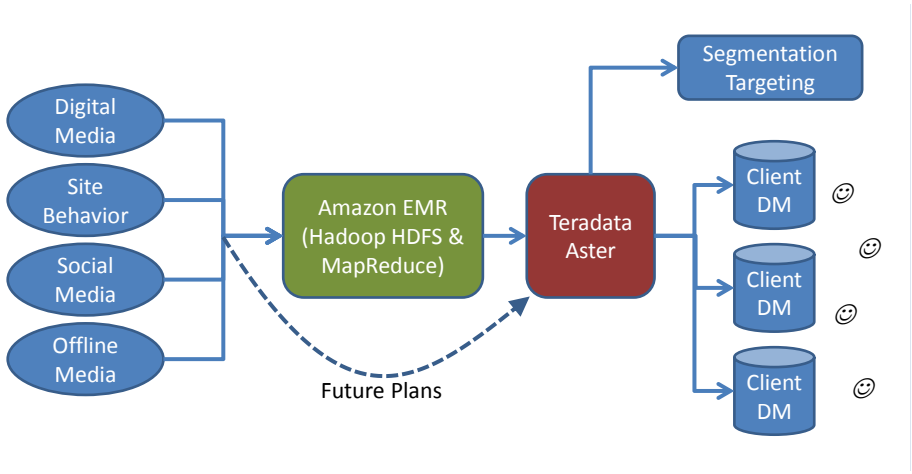


Figure 3 - Architecture for Company B.

Teradata Aster

The Teradata Aster platform provides an analytical library of SQL-MapReduce functions that we use as building blocks for our processing. Over the long term, we are consolidating the Amazon EMR jobs onto the Aster platform to minimize the data hops. The less we touch the data to get it into its final form and to its final resting place, the more efficient we will be.

The focus of the Aster platform is to manage the media information at the aggregate cookie level so that we can see who is consuming what media. At the moment, we are 'kicking the

tires' to learn the capabilities of the Aster platform, with the intent to consolidate other processing onto Aster.

The segmentation analytics use the Aster SQL-MapReduce functions for sessionization and pathing, along with our custom matching algorithm to match session identifiers and create a centralized identifier for the customer. This processing is performed with a single pass of the data, which is more efficient than normal SQL processing. Depending on the client, specific data about the client's business is brought into their data mart because of their unique data needs.

Core Competences

As a core competence, Company B knows all about digital marketing and the technology driving digital marketing. Our value to clients is to question clients about whether they have looked at this technology or that technology. And, if we have successfully applied a new technology to a client's problem, we then would suggest to other clients to look at that same technology.

We focus on our core competence in digital marketing optimization. If we need resources outside our core competence, we will purchase it outside. In fact, we have leveraged offshore resources to reduce our costs. However, if it is part of our core competence, then we will build or acquire it internally. The build-versus-buy decisions with a software project are often judged from this perspective.

He gave the following example:

We could have built those analytic components in the Aster platform. However, I also know that developing those components is not our core competence, so we purchase them from Teradata Aster. However, using those components in unique ways to satisfy our clients' needs, this is our core competence.

Leveraging Open-Source Software

There is pressure to leverage open-source software, like Apache Hadoop. Under the right circumstances, open source software is very attractive. As this technology matures and more people use it in new ways, the packaged software vendors will be pressured to extend their functionality to maintain their competitive advantage over open-source alternatives.

Aster SQL-H Connector

To leverage open-source software, Teradata Aster offers a connection layer, called Aster SQL-H™, to the Hadoop Distributed File System (HDFS) data based on HCatalog meta-data. This feature enables HDFS files to be read as native Aster tables, thus leveraging the Aster Database analytic functions. Although the intent is to move daily processing to the Aster platform, the company will continue to store and manage some data in HDFS, probably using SQL-H™. Since Teradata Aster charges by the data volume managed by the Aster platform, the company feels that it needs the flexibility to utilize both options.

The current data retention policy is at least 12 months. Since more than five terabytes are acquired per day, more than one petabyte of data is being stored in Amazon EC2 every year.

The company is working on a long-term retention policy to determine the business value of archiving years of acquired data.

Lessons Learned

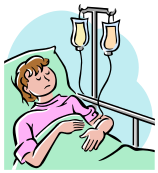
He shared the lessons learned are about the ups and downs of adopting new cutting-edge technologies.

"Being aggressive with new technologies is risky since there are often unexpected obstacles. Various features and functions that you assume with traditional technology may not be present in new technology. Surprise! You do pay for being on the bleeding edge!"

Summary

As an ad agency for digital media, Company B uses analytics to understand the impacts that digital media is having on the marketplace of their clients and to optimize the 'spend' for that digital media. The analytics for their multi-attribution model requires large volumes of continuous data from numerous sources. Custom data marts for each client manage the results and support internal teams in providing services to their clients. The Aster platform provides packaged analytics for agile analyses and a way of consolidating core data "to minimize hops." Apache Hadoop will continue to provide low-cost storage and filtration of raw data. The use of Aster SQL-H™ is being considered as a means of increasing their data retention period to greater than 12 months.

Company C – Large Healthcare Provider



Company C is a large healthcare provider with dozens of hospitals, many medical clinics, and thousands of caregivers. As a strategic IT initiative, they are serious about handling their big data properly so that managers can access detailed operational data to improve healthcare delivery, to reduce unnecessary expenses and minimize medical liabilities.

Background

The Director of Business Intelligence was interviewed. He described their experience with analytical platforms. After considering several options, they decided to invest in the Teradata Aster Database platform for reporting. By consolidating all their data marts onto this one platform, the overhead of loading and transforming data from diverse sources and deploying query tools to diverse users is simplified, thereby reducing costs. With 18 months of production experience with the new platform, the company is "*tickled pink with the performance.*"

Potential of Big Data

The director described healthcare as a business that generates voluminous detailed operational data, which is usually not retained for more than a few days. The Electronic Health Record system tracks the minute-by-minute event flow of medications and procedures given to patients. Most of this information comes from specialized medical equipment that generates real-time data, such as an MRI or CAT scan. Although some images are retained within the

Picture Archival System (PAS), most of the data stored on medical devices soon "*drops on the floor*" (i.e., deleted to accommodate new data).

An important use case is managing patients' notes created by physicians. The physicians submit their medical reports as voice recordings, which are transcribed into unstructured text. These reports contain very important information about the patient, such as future diagnoses. The physician is "extremely candid about the patient, more than anywhere else." There has been work on designing dictation templates into which the unstructured text could be mapped during transcription. However, this approach is a partial solution since the templates cannot differentiate among the diversity of medical situations. Long discussions are often placed in a single cell of the template. Text analysis could provide more structure and hence more value to the physician's report.

To better understand the dynamics and economies of their caregiving business, medical analysts would "*like to get their hands on that data.*" The only way to access this detailed data now is to retrieve the record and extract the appropriate data – all manually.

Future Plans

As a future project, Company C is planning to develop a Hadoop platform to eventually collect as much of this unstructured data as is possible. It will be stored as raw unstructured data in the Hadoop Distributed File System (HDFS), as shown in Figure 4.

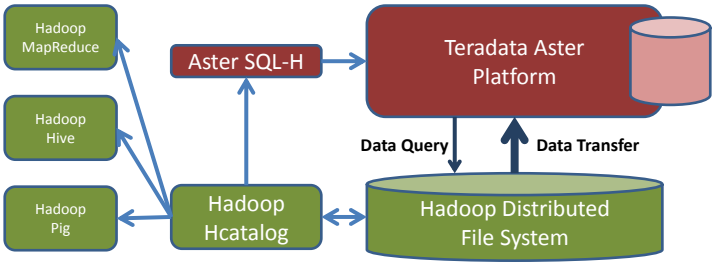


Figure 4 - Aster SQL-H™ Connector to Hadoop HDFS.

There are various methods working with the HDFS data, all of which use the Apache HCatalog table/storage management to share schema information for interoperability. First, the Hadoop MapReduce engine supports job submission and job scheduling as part of the Apache Hadoop package. Second, Apache Hive supports data aggregation functions with its HiveQL language, which are compiled and submitted as MapReduce jobs. Third, Apache Pig creates MapReduce programs with a higher-level procedural language. Finally, Teradata Aster supports an extension to their declarative SQL-MapReduce language called Aster SQL-H™, which accesses the Hadoop Hcatalog and directly joins only the data required. Hence, business analysts can work with Hadoop data like another flat table through standard ANSI SQL and various BI query/reporting tools.

Aster SQL-H to Access Hadoop Data

He remarked that SQL-H™ is being considered as the link between the Hadoop platform into the Aster platform since the two environments are remarkably complementary. Teradata Aster

provides a managed MPP analytic discovery platform that blends SQL and MapReduce, while Hadoop is a low-cost open-source platform that also uses MapReduce. Aster provides SQL-MapReduce to integrate SQL and MapReduce on the Aster platform and SQL-H™ to integrate SQL with Hadoop HDFS data.

The BI director realized that SQL-H™ enables his development staff to "*remain with the SQL*" paradigm and avoid "learning and operating another system to access this data." He could also "*leverage the current SQL skills to analyze Hadoop data.*" The developer would need to know the syntax and functional difference in SQL-H™, but he judged that this is "*not that far of a stretch.*" He estimated that the ROI on this project would be greater since all the data in Hadoop is accessible using their existing analytics environment. Thus, he surmised that this project had "*a clean business use case*" with minimal risk in both financial investment and resource constraints. We could "look at data that we don't look at today" with a quick development effort. "*At the end of the day, we have to survive on adding value*" for the company.

When asked about the cost of Hadoop storage, he estimated that it would be considerably less because of the use of thousands of low-cost fault-tolerant commodity servers to handle data volumes in the petabyte range. By using SQL-H™ to filter the Hadoop data, only the result set, which should be in the gigabyte range, would be pushed back to the Aster platform. He concluded that, "*Only the actionable data should be managed by the Aster platform.*"

He described their status as "*in the thinking stage*" with no SQL-H™ activity…"*yet, but we are certainly looking into it*" as it is "*on our technology roadmap.*" He finished with, "*If we could provide this data access, we would win the ballgame!*"

Summary

Company C is a large healthcare provider that is seriously exploring ways of managing and exploiting detailed operational data. Over eighteen months, the company consolidated various data marts into the Teradata Aster platform. If the voluminous operational data could be economically acquired and stored, the analytics of the Aster platform (such as text analysis on patient notes by physicians) has the potential to provide insights into improving their healthcare delivery. In particular, they are currently investigating the Aster SQL-H™ Connector as a channel to large data sets stored under HDFS.

Tradeoffs

In planning analytical architectures, companies should consider the following tradeoffs, such as scalability, time-to-value, skill inventory, and multiple data channels.

Monetizing Time-To-Value

A critical tradeoff in making investments in analytical architectures is monetizing the time-to-value factor. In other words, what is the business value that results from faster 'discovery' cycles? Analytical processing often consists of an analyst who is continually going through

cycles of collecting, refining, modeling, and testing data. It may take many such cycles before the analyst produces valid and practical results for the business. With current technology, it is possible to reduce those cycle times from days to minutes, thus increasing the productivity of the analyst by a hundred-fold. Further, these faster discovery cycles can enable the analyst to work as fast as they can think, permitting a high level of uninterrupted concentration on the analysis problem.

How should the company monetize the productivity increase of the analyst? In the past, executives would decide that the analyst could take a week or two and avoid the additional expense. In the present, many companies live and die based on leveraging 'now' information that guides business processes, minute by minute, as they unfold. Several years ago, a major retail company boasted that their point-of-sale data was in their data warehouse by the time the shopper left the parking lot. That same company considers that boast so past season and is seeking ways of up selling to their customers as they wander the store floor, long before they leave the parking lot.

Planning for Scalability

A critical tradeoff in making investments in analytical architectures is setting the priority on scalability. In past decades, only the big companies worried about IT investments in equipment purchases, data center cooling, and the like. Small companies did without. Now many start-up companies are exploiting analytics to open new business niches having explosive growth. From day one, those small innovative companies must think deeply about scaling their IT capabilities to service big clients, who internally are unable to exploit the same analytics. Hence, rapid scalability of analytical architectures – petabyte data volumes, sub-second operational analytics, thousands of concurrent users -- are hard requirements, not lofty dreams. By contrasting analytical requirements as current versus future, the company can assess the priority of investing in scalability.

Assess and Manage Skill Inventory

Much of the technology surrounding big data analytics has evolved in recent years. Hence, persons (either employees or contractors) with the proper skills are often in short supply. Further, the technology is changing fast so that acquiring new skills should be a continual learning process.

Juggling Multiple Channels

The initial applications of big data analytics acquired, cleansed, filtered and transformed one data source, such as a Twitter text stream. As analytical applications, like marketing optimization, became more sophisticated, it was apparent that many data channels were required for analytics on the total customer experience involving multiple touch points (like the corporate website, Twitter, Facebook and others).

The manual effort to acquire, cleanse, filter and transform one data channel is huge. Multiple channels require N times 'huge' effort plus… generating cross-channel linkages so that all the touch points for a specific customer are interrelated.

Insights

This section summarizes and extracts practical insights for professionals pursuing big data analytics within their companies.

Maturing into Enterprise Architectures

The maturing phase that big data analytics is undergoing currently is similar to the evolution of data warehousing from a departmental scope into enterprise architectures. Even today, there are certainly many reasons for a department DW, such as a clear business objective, restricted user base, and single functional emphasis. However, we have learned over the decades that there are huge business motivations to extending and integrating departmental DWs into an enterprise DW, such as cross-functional business opportunities, pervasive support of users (even customers), and operationalizing analytics into business processes.

The same is currently happening to big data analytics. This implies that there will continue to be good business reasons for isolated, single-purpose analytics projects. However, as soon as the business viability of those projects is established, the debate of how to integrate that functionality into the enterprise architecture should start with vigor. The company cannot ignore the tough issues, such as security, data governance, privacy, and scalability, without curtailing future business potential.

Data ownership is one of these tough issues that has traditionally been a tug-of-war between business users and IT managers. IT wants to own all the 'official' data about the corporation, while business users want to use that data with flexibility and responsiveness.

Clash of Tech Cultures

There is considerable debate about whether emerging technical alternatives, such as procedural versus set operations, no-SQL versus SQL, and the like, will dominate over their older cousins. History clearly shows that tech waves never obviate previous practices. After the initial hype subsides, the savvy professionals find various optima in the blending of those tech waves. Over time, best practices emerge from continuous incremental integration of new technologies into coherent architectures.

The same is happening with big data analytics and was especially apparent in the companies interviewed. With any initial proof-of-concept project, the driver is to get the job done, as soon as possible, as cheaply as possible, with whatever tool and skill is available. Once business viability is established, the driver should shift to determining the best way to accomplish the job. Hence, a company may need to acquire new tools, new skills, and a whole new technology to properly satisfy the required objectives. The point is… Do not constrain your futures by choosing between A or B. Be knowledgeable about both A and B. And constantly search for the complementary blending A and B.

For example, an early issue for this study was whether Apache Hadoop or Teradata Aster was better. The interviews quickly illustrated that this issue was a fluid comparison that changed over time. The companies were influenced in their technology decisions as several factors, such

as prior skill sets, depth of analytics, frequency of discovery cycles, and support of operational business processes. The pivotal change was the introduction by Teradata of the Aster SQL-H™ connector, which allowed Aster Analytics to access directly HDFS data. The debate shifted from choosing between A or B to minimizing data movement and thus decreasing the time to perform analyses.

Moderate TCO Debates

In the first section on Challenges, the point about TCO was that executives should think strategically and broadly about assessing analytical architectures. One way of justifying this statement is to review the business objectives for big data analytics for the three companies.

Company A	E-commerce retailer that provides customers with a customized online shopping experience by being more responsive to customers' behavior. Hence, the company needs to acquire cyber-data about all touch points with the customer.
Company B	Digital-media ad agency that optimizes all digital marketing for clients. Hence, the company acquires data on all customer touch points across all media channels related to their client.
Company C	Healthcare provider that improves healthcare delivery and reduces medical expenses. Hence, the company acquires data on medications and procedures for patients and wants to expand data collection to all medical devices and from patient notes by physicians.

Note that these business objectives strike close to the core business competence of the companies. Could the company continue in business if their analytical applications cease to operate properly? For one day? For one week? These questions paint the nature of a TCO assessment in very practical and realistic terms. Therefore, TCO debates within a company should be moderated by skilled executives knowledgeable about the business reasons for creating and maintaining analytical applications.

Riding the Tech Waves

We have seen many tech waves over the past decades. However, the tech wave behind big data analytics is obviously one of the more challenging and volatile. To understand the implications and trends with the many disciplines comprising big data analytics requires a unique individual. Companies that are making substantial investments in analytical architectures must be knowledgeable of these implications and trends. The technology is moving very fast. Companies must satisfy current requirements while constantly planning for future requirements. Therefore, companies require the skills of a competent CTO-like person to ride this tech wave and to mentor others to do the same.

- - - - -

In summary, any company that incorporates big data analytics into its business processes should face and resolve these questions:

1. Will you be able to support the complete analytical value chain from data to action, thus achieving tangible business results?
2. What is the plan to mature the current/proposed analytical applications across the scope of the enterprise?

3. How will you leverage the best from various tech cultures to produce the desired business results?
4. Do you have a constructive way of debating the TCO for new innovative analytical applications?
5. Do you have a good surfing instructor so that you can ride the volatile tech wave of analytics?

Endnotes

[1] "The discovery and communication of meaningful patterns in data" is the definition in Wikipedia at http://en.wikipedia.org/w/index.php?title=Analytics&oldid=510946342 retrieved 6 September 2012. It is a weak entry but covers the basics.

[2] As originally defined by Doug Laney in 2001. For more background, see http://en.wikipedia.org/wiki/Big_data retrieved 6 September 2012. It is a good entry with many examples.

[3] For background on Apache Hadoop, see http://en.wikipedia.org/w/index.php?title=Apache_Hadoop&oldid=510910492 retrieved 7 September 2012. It is a good overview of the entire Hadoop ecosystem and community. Note the funky component names! This is changing on a monthly basis, driven by Google, Yahoo, Amazon, IBM, Cloudera, Hortonworks and others.

[4] For background about Teradata Aster, see the Resource section at http://www.asterdata.com/, which has good whitepapers, analyst reports, webcasts and more.

[5] This value chain was adapted from a presentation by Mayank Bawa of Teradata Aster. It is also a tip of the hat to *The Corporate Information Factory* by Inmon, Imhoff & Sousa, Second Edition, 2001. http://www.amazon.com/Corporate-Information-Factory-W-Inmon/dp/0471399612

[6] http://en.wikipedia.org/wiki/Random_forest

About the Methodology

The methodology of this study is to listen carefully to several pioneering companies in the big data analytics area. The intent is to contribute to professional education—to share the insights with other IT professionals so that we can mature as an industry, amid escalating business challenges and rapidly evolving technology.

The sample of Teradata customers was small so these conclusions are tenuous but insightful. By leveraging the open access to Teradata customers, we have a glimpse into the complex issues involved. Based on the quality of the interviews, this sample is representative of the issues and trends in this emerging area.

The primary author is Richard Hackathorn of Bolder Technology with substantive contributions from several Teradata colleagues: Steve Wooledge, Manan Goel, Mayank Bawa, and Kevin Pratt. We are appreciative of the companies and professionals who were willing to share openly their experiences. Finally, we are appreciative of Teradata Corporation for their assistance and sponsorship of this study.

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 IBM Gold Consultants and the Boulder BI Brain Trust. 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 Sponsor

Teradata is the world's largest company focused on analytic data solutions through integrated data warehousing, big data analytics, and business applications. Only Teradata gives organizations the advantage to transform data across the organization into actionable insights empowering leaders to think boldly and act decisively for the best decisions possible.

EB-7471 > 0113