Paper 1828-2014

Integrated Big Data: Hadoop + DBMS + Discovery for SAS® High Performance Analytics

John Cunningham, Teradata Corporation, Danville, CA

ABSTRACT

SAS High Performance Analytics (HPA) is a significant step forward in the area of high-speed, analytic processing, in a scalable clustered environment. However, Big Data problems generally come with data from lots of data sources, at varying levels of maturity. Teradata's innovative Unified Data Architecture (UDA) represents a significant improvement in the way that large companies can think about Enterprise ever, Data Management, including the Teradata Database, Hortonworks Hadoop, and Aster Data Discovery platform in a seamless integrated platform. Together, the two platforms provide business users, analysts, and data scientists with the ideally suited data management platforms, targeted specifically to their analytic needs, based upon analytic use cases, managed in a single integrated enterprise data management environment. The paper will focus on how several companies' today are using Teradata's Integrated Hardware and Software UDA Platform to manage a single enterprise analytic environment, fight the ongoing proliferation of analytic data marts, and speed their operational analytic processes.

INTRODUCING SAS HIGH PERFORMANCE ANALYTICS

SAS In-Memory architecture has been a significant step forward in the area of high-speed, analytic processing for Big Data. One of the major themes of new SAS offerings over the last couple of years has been High Performance Analytics, largely using In-Memory clustered technology to provide very fast analytic services for very large datasets. Together, both SAS Visual Analytics and SAS Visual Analytics have made significant strides to affordable analytic processing, using largely a very similar in-memory architecture.

Key to SAS HPA and SAS VA is clustered processing, also known as Massively Parallel Processing (MPP). This model enables SAS deployments to scale cluster size to support larger data, higher user concurrency, and greater parallel processing. However, the most significant benefit for SAS users is blazing speed, and both environments high speed memory centric techniques to achieve extremely fast analytic processing.

For example, SAS HPA reduced the analytic processing time for one Teradata customer down from 16 hours for a specific analytic process to 83 seconds. This improvement in speed was dramatic – everyone likes to get done quicker. But the biggest impact for this company was that it now enabled their users to "experiment more", try more advanced model development techniques, and utilize the mantra of "failing faster" – if it only takes a couple of minutes to fail, you're much more willing to try again with another variant on the process, as opposed to if the overall process took a week.

For SAS VA users, same story – enormous 1 billion row analytic datasets can be pulled up and visualized in a couple of seconds. Users can apply different analytic techniques, slice and filter their data, update different visualization techniques, all with near instantaneous response time.

The SAS In-Memory architecture comes in two data management flavors – one that utilizes a MPP database management platform for storage of all of the data, and another that utilizes a Hadoop file system cluster for persistence. Both work, both are fast, both can scale when, especially when working with advanced clustered Massively Parallel Processing (MPP) style database models. However, for many complex organizations undertaking large scale "Big Data" projects, neither is a "one size fits all" solution.

The most economical model utilizes Hadoop, running HPA directly against data scattered across the nodes of a distributed file system. With this model, relatively low cost servers can be used to support the Hadoop workers, connected together by high speed networking, with processing managed via the MapReduce. This distributed process gives the same benefits of many of the MPP model databases, but with fewer bells and whistles, at a lower cost.

IF HADOOP IS THE CHEAPEST, THEN THAT IS THE RIGHT PLATFORM FOR MY ANALYTICS, RIGHT??

Maybe. If you have massive data files, and want to make them available for analysis, without concern for data structure, then storing them on a economical file system may be the best way to persist them. If you are continually gathering terabytes of log data on a daily basis, some of which may not be interesting or even relevant, then persisting that data in Hadoop can be the most effective approach. If you don't care about providing granular level of security control, and are prepared to have dedicated Linux server administrator manage the cluster as it grows –then sure. If you code in Java, or a comfortable with writing shell scripts to accomplish basic activities, then maybe. If

you don't care about accessibility, or whether you want to make the data available via business intelligence tools for business users, then Yes. If there are just a few analysts on your team, and analytic concurrency is not a requirement… Or if high availability is not a concern, and neither is end to end fault tolerance… If you don't care about cost for bringing in trained, specialized resources to support your projects, and don't care that that Hadoop specialists can be very expensive to procure.

If you answered Yes to All or Most of these questions, then Hadoop may likely be the best platform for you for your analytic use case. For an individual data scientist, already versed in Linux administration, already familiar with Java coding techniques, then a Hadoop Data Mart may be the ideal platform. In this case, the data scientist can gather data from dozens of other data sources, copy it over to their Hadoop distribute file system, and leverage the SAS analytic toolset to explore at high speed. They have a lot of flexibility in this environment, and with free reign to structure the data as they see fit.

However, for most companies, the above Hadoop-appropriate use cases are generally only appropriate for certain classes of applications, certain types of data. For many companies, there fit may be very specific classes of problems that align well to the prerequisites, and for these specific cases, leveraging a Hadoop cluster will likely be the most cost effective solution to the problem.

TURNING ONTO THE HADOOP "DATA CUL-DE-SAC".

With big data projects, many companies have recently started to experiment with analytics projects spun up on Hadoop clusters, used by a couple of up and coming data scientists, who fit the mold of the ideal Hadoop centric advanced users identified above. In some cases, we started to refer to the projects as "Data Cul-De-Sacs" – where it's easy to turn down the road to start the project, but it may be difficult to get back out with an integrated solution. Frequently, the experiment that drives a Hadoop cluster are highly specialized business problems, so it's simple to turn down the street to making a redundant copy of the data, perform value added analytic services, and potentially provide a high value analytic answer. However, for many of these projects, once you've gone down this road, it ends – and the data doesn't easily come back out again. For cases where the answers are highly valuable, they may be accessed by business users via SAS Visual Analytics or another similar platform, but the result is the dreaded "dependent data mart"

The challenges include:

- Copying data onto this disconnected Hadoop cluster can be a slow, tedious process. Worse yet, copy high value answers back to some other operational system for tightly integrated services can be equally as painful. If you're a highly valued data scientist, you may spent a large amount of your time waiting for large datasets to be copied – just to get started on the analytic work that matters.
- If your experiment is successful, once you spin one of these dependent Hadoop clusters up, the cluster needs to be managed. That likely requires ongoing Hadoop administration, managed ETL from desperate data sources, accessibility control, etc.
- Also, by definition, once you copy data to a redundant data mart, it's redundant! This means that the same data is in two different places, and is being transformed differently in each – which ultimately will lead to varying results.

In the end, Teradata strongly advocates against specialized "Data Cul-De-Sacs" – there are far better benefits to be gained through tight integration with other data management platforms.

TERADATA UNIFIED DATA ARCHITECTURE – PURPOSE BUILT DATA MANAGEMENT.

Teradata has long been the industry leader in large scale Enterprise Data Warehouse platform, and most of the biggest companies, with the biggest data warehouses, are running on Teradata . The Teradata Unified Data Architecture is an evolution of the model, and recognizes that an EDW isn't always the right place for certain classes of data. For example, while Teradata Data Warehouse scales well to support complex data models, access from hundreds of users, with varying workloads, in a highly managed, highly governed, rigorously controlled environment, there are classes of data that are likely not going to benefit from Teradata services.

For example, while Teradata may support certain unstructured data, log files, or streaming sensor outputs to be written into the Teradata database in the form of Binary Large Objects, putting them there may not provide any analytic lift to the data scientist. Data types like BLOBS are very capable of store virtually any random data content, but if the end goal is to support parsing, text analytics, keyword lookups, then other stored methods may be preferable.

For that reason, we introduce the Teradata Unified Data Architecture.

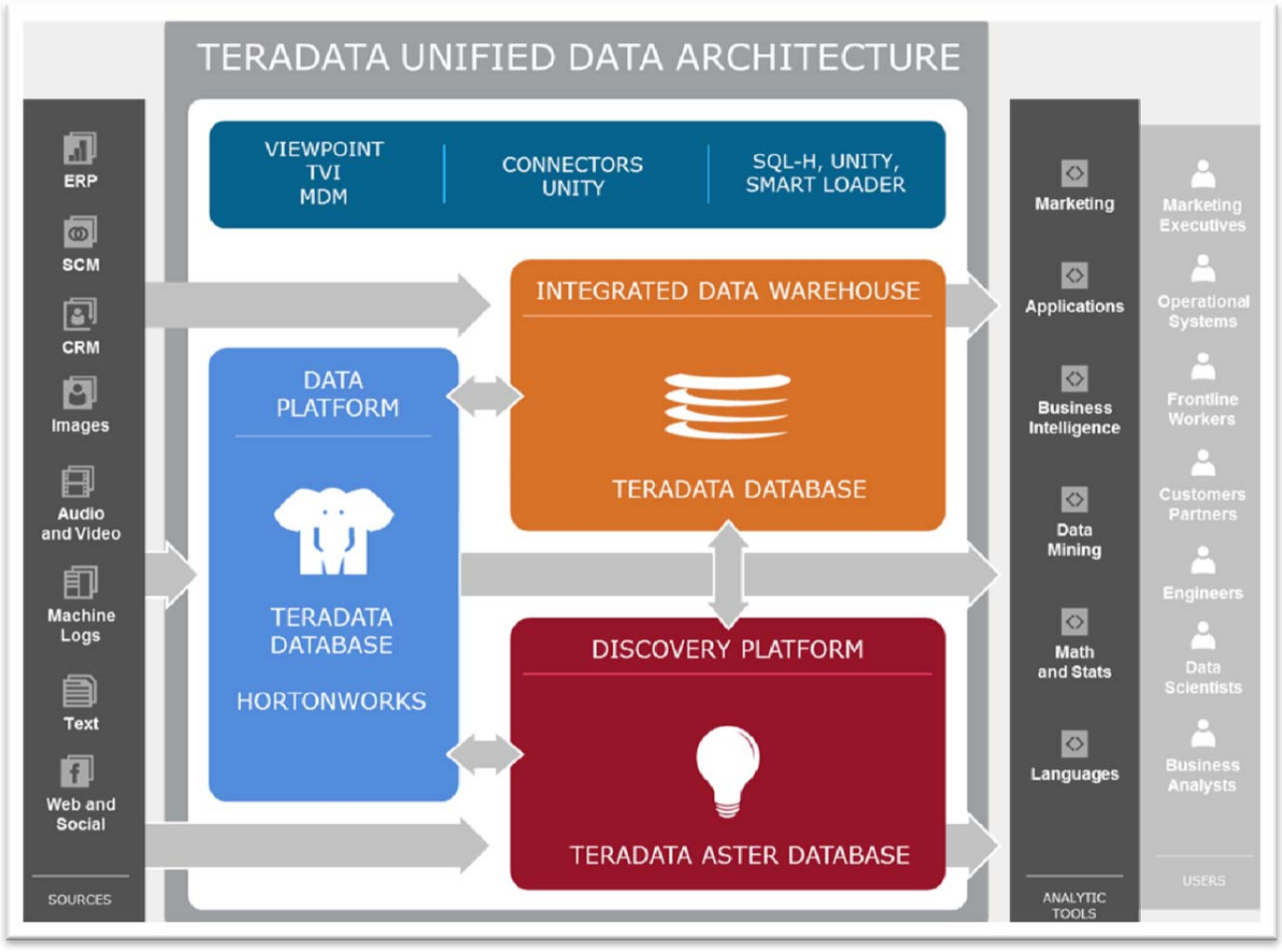

Figure 1: Teradata Unified Data Architecture

The Teradata Unified Data Architecture provides 3 distinct, purpose build data management platforms, each integrated with the others, intended with specialized needs:

- Enterprise Data Warehouse - Teradata Database is the market-leading platform for delivering strategic and operational analytics throughout your organization, so users from across the company can access a single source of consistent, centralized, integrated data.

- Teradata Discovery Platform - Aster SQL-MapReduce delivers data discovery through iterative analytics against both structured and complex multi-structured data, to the broad majority of your business users. Pre-packaged analytics allow businesses to quickly start their data-driven discovery model, that can provide analytic lift to the SAS Analytics Platform.
- Data Capture and Staging Platform – Teradata uses Hortonworks Hadoop, an open source Hadoop solution to support highly flexible data capture and staging. With Hortonworks, Teradata has integrated it with a robust tools for system management, data access, and one-stop support for all Teradata products. Hadoop provides low-cost storage and pre-processing of large volumes of data, both structured and file-based.

With these 3 elements, the UDA provides flexibility to SAS HPA users, enabling them to manage their data in a specific purpose platform, directed by the specific requirements of the analytic solution involved. This flexibility is critical for SAS HPA users that are looking to integrate their analytic processing with enterprise data access services, that are looking to avoid the "Data Cul-De-Sac", and integrated their analytic directives into other corporate decision making.

For most Teradata customers, they already utilize a Teradata EDW to gather data from multiple operational data sources. Integrated together, to provide a "single version of the truth" that can be interfaces from hundreds of downstream applications, or thousands of business users. However, this data is generally relational, organized into Tables, Rows and Columns, accessible by SQL. Depending on the data maturity that one may be working with, other platforms may be better suited:

- The Enterprise Data Warehouse is targeted at rigorous, highly structured, "ready for prime time" data that can be used by business and analytic users around the enterprise. EDW's typically also are factoring enterprise data readiness concerns, not only with the ability to scale to support more users, wider range of user applications, or larger volumes of data, but also scale to anticipate potential failures, keep running in the state of problems, and provide management and monitoring environment to ensure that the environment is continually to support data access from around the business. It is ideally suited for modeled and integrated data to support end-to-end optimization and operationalization
- The Discovery Platform is intended to for semi-structured data, still in need of analytic processing, but organized into functional areas sufficient that it can fit into a table oriented data structure. In this case, there's a tradeoff, where we are dropping some of the user concurrency requirements, in exchange for high levels of analytic flexibility,
- The Hadoop Data Staging platform is intended to be an economical platform to capture and store varying types, the data that hasn't yet been fully structured, and doesn't yet need to be accessed for operational reporting across the enterprise. Frequently, it's used to collect sets of enormous files, like web logs, machine sensor outputs, or even web log outputs, all of which have some analytic value to them, but most importantly just need to be stored in a large clustered file system.

Depending on the nature of your analytic case, you may only need one platform to address the specific need. For example, if enormous web log analysis is the primary purpose the analytic solution, then running SAS VA on Hadoop may be the ideal platform. However, if you also require operational data, it would be best to leverage both your EDW and Hadoop together. If the same data may be used to drive iterative analytic processes, then also including an analytic data mart platform like Aster may be ideal.

For some large companies, they utilize the different aspects of the UDA side by side, each with special purpose perspective on the end to end data lifecycle, evolving and maturing data from its "staged" landing area on Hadoop, right up to the "ready for action" enterprise data warehouse environment.

TERADATA PLATFORM FAMILY FOR UDA + WHERE SAS HPA FITS IN

To that end, Teradata offers a complete platform family, specifically targeted at all of the specialized needs of the specialized class of analytic users.

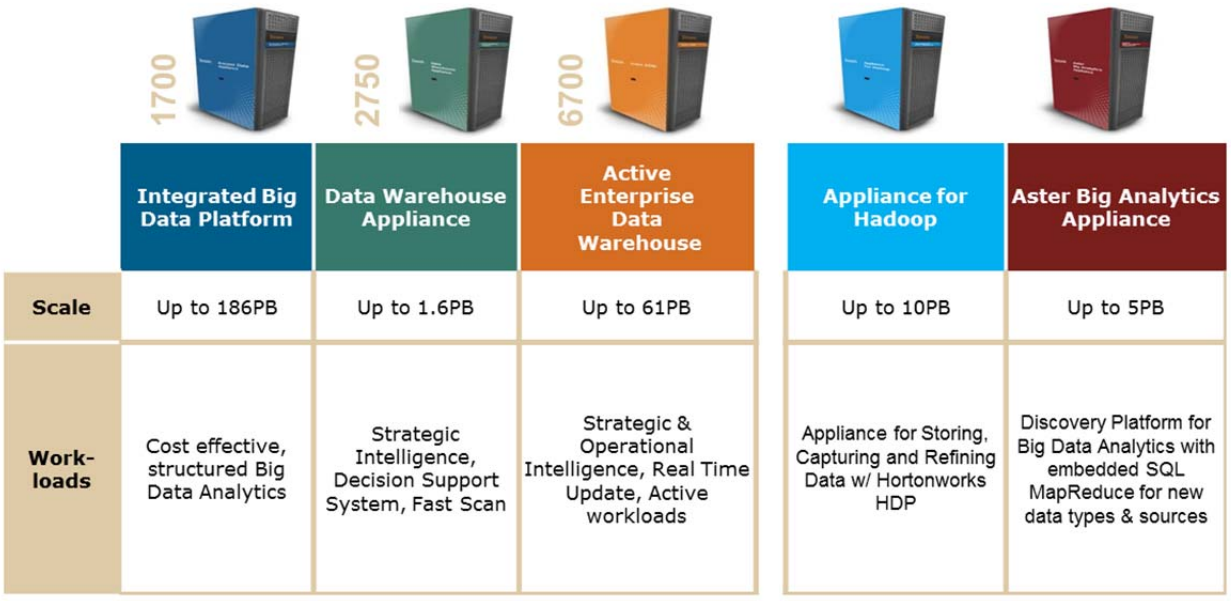

Figure 2: Teradata Platform Family

Depending on the size of the data the company is working with, the number of users, the intended response rates for queries, Teradata will direct companies towards different size Data Warehouse Appliances. Additionally, as companies requirements expand into less structured Big Data, either the Appliance for Hadoop, or the Aster Big Data Appliance is ideally suited to integrate directly into the overall Teradata Complex. In the Unified Data Architecture, all of these appliances share a common data management backbone, meaning that each of the clustered appliances can quickly pass data between systems, faster than via a traditional BI interface.

THE TERADATA APPLIANCE FOR SAS HIGH PERFORMANCE ANALYTICS

Integrating side by side with the rest of the Teradata Platform Family, the Teradata Appliance for SAS High Performance Analytics extends the analytic capabilities of the Teradata environment, enabling SAS In-Memory architecture directly onto the Teradata environment. With the HPA Appliance, Teradata extends any UDA data platform with new HPA capabilities, conceptually like adding as “SAS Math Nodes” into the Teradata database. These Math Nodes support the specialized SAS MPP architecture, providing a computational environment for SAS to run on, that doesn’t negatively impact the database platform that is also running on the same physical systems.

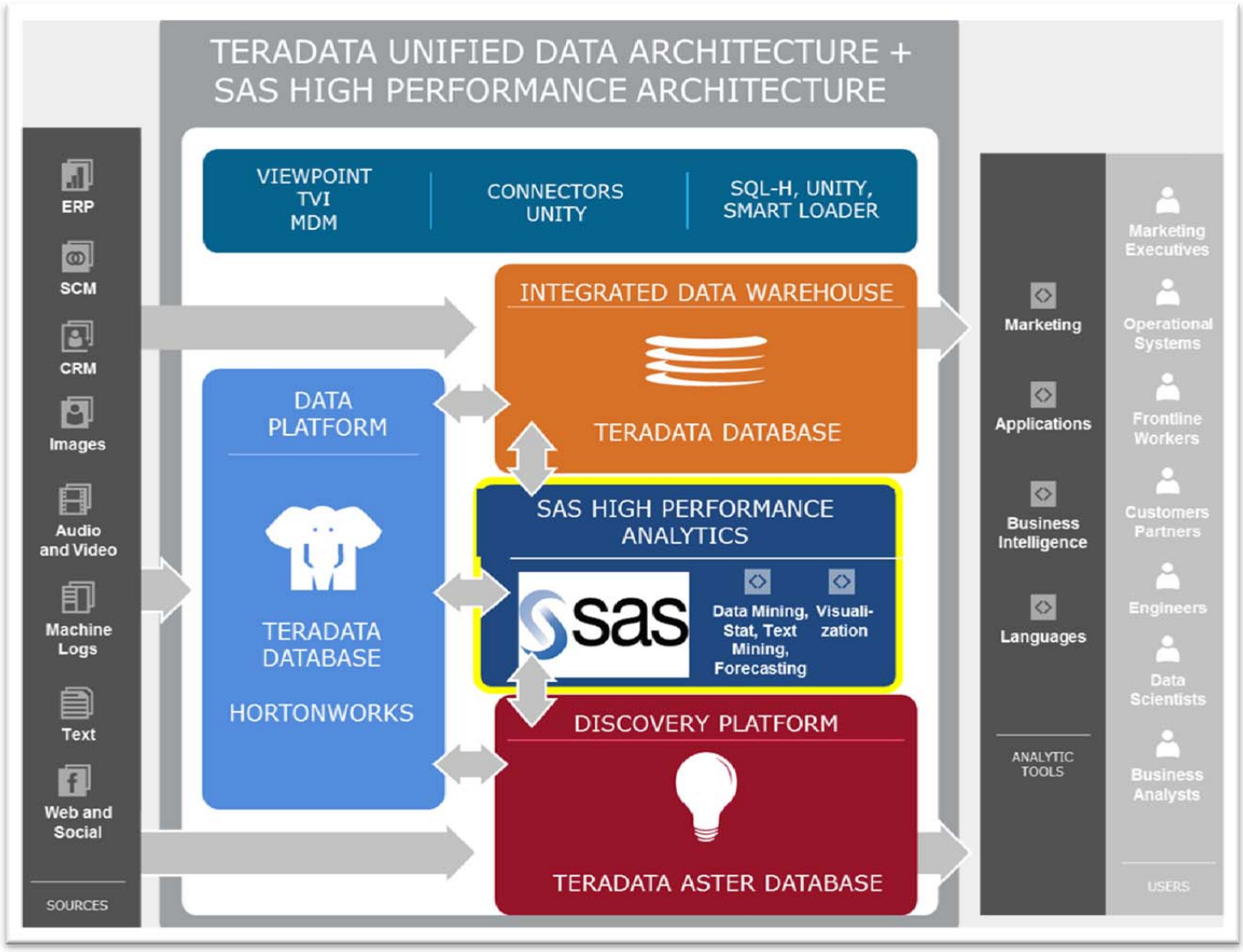

Figure 3: UDA W/ integrated SAS HPA

What makes this special, however, is the Teradata Connect Model. The Teradata Appliance for SAS HPA utilizes Teradata Connect as a special high speed parallel data passing architecture, allowing dozens / hundreds of parallel pipes of communication to occur between the different large MPP systems. Unlike traditional BI tools, which can be relatively slow because they are limited by only a single "pipe" from a Teradata data source, the SAS HPA architecture essentially has each node of the Teradata SAS Appliance act as a parallel reader of data, read just what it needs, local from that node, from another specific MPP node of the data source, dividing the overall data stream into hundreds of parallel pipes, and then leveraging the high bandwidth capabilities of the Teradata BYNET / Infiniband backbone. The result is extremely fast data spooling architecture that gets data to the Data Mining, Text Mining, Forecasting and Visualization very fast, starting those processes in a few seconds, as opposed to the dreaded wait encountered over a traditional public network.

FABRIC BASED COMPUTING

As mentioned before, Massively Parallel Processing systems can process large amounts of data very efficiently, across the interconnected nodes of a private network. These MPP queries take advantage of the distributed nature of analytic data, push processing to each individual MPP node where appropriate, and spooling data between nodes when required. When an external tool needs to interface with that data via a query, such as SQL, they query processor isolates the processing to the specific node(s) that has the data, spool multiple results back to a single gathering node that returns the result set stream. Atomically, this process is very efficient, and scales well as data get bigger.

However, if you want to connect two MPP systems together, the results may likely be sub-optimal. For example, for most vendors, if you were to attached an MPP data warehouse to an MPP Hadoop cluster, and attempt to copy a large table from one to the other, this process is likely not optimal. In many cases, the SQL processing on the Data Warehouse may occur, but then stream the resultant data as a single pipe (or as a few parallel threads using load utilities) over a slow 1GbE public network connection, as a sequential stream of data. When the data reaches the Hadoop side, the data stream would then be re-distributed out to the varying nodes of the Hadoop cluster, a process that itself is likely very efficient. However, there is a bottleneck is in the middle, and copy a terabyte table from one to another may take many hours to complete.

The Teradata Unified Data Architecture leverages the concept of Fabric Based Computing, where it utilizes redundant high-bandwidth networks to connect all of nodes of each MPP system to each other. All of the different platforms on Unified Data Architecture all leverage similar dedicated network multi-node

In the world of Fabric Based computing, Teradata integrates multiple MPP systems into a complex, multiple cabinet environments, each all sitting on a single common high bandwidth Infiniband backbone. For traditional query's, each MPP database appliance can operate atomically, and to service its users with their own queries in a high efficient fashion. However, for Fabric Based Computing enables any node in any MPP systems to communicate with any other connected node. For tightly integrated, sophisticated applications like SAS HPA, this means that any node of the Teradata 720 SAS HPA Appliance could read / write data to any of the database nodes on any of the appliances, and to achieve ultra-high speed data interchange with potentially all of them at the same time.

Since the Teradata Appliance for SAS High Performance Analytics sits on the same network fabric, it can work with the Teradata and Hadoop data sources in the highly efficient parallelized way. Data from each can be quickly loaded to the SAS HPA processing engine for rapid model building, or loaded into the SAS VA memory pool for rapid visualization, the Teradata Appliance for SAS HPA can act as the "Math Nodes" for multiple database systems, all at the same time. Individual queries can be processed souring raw data from a Hadoop staging platform, pushing the lightly process output result set to a Teradata system. SAS Visual Analyzer, which utilizes the same In-Memory architecture, can help perform Data Discovery on the raw data in the Hadoop instance, visualized side by side with the mature data from the Teradata EDW.

EXAMPLE - END TO END BIG DATA ON UDA WITH SAS HPA

At a high level, let's follow a sample scenario through the end to end data management life cycle that leverages SAS analytics on top of Teradata UDA. In this example, we'll focus on the common Big Data use case of Web User Behavior Analysis via Web Logs, and using that providing meaningful processing to users' of a contact management system.

Web User Behavior Analysis

Step 1: Gathering

Web Logs are typically the focus of Big Data Analytic problems – Web Logs themselves are fairly boring pieces of data, and include lots of data elements that today are not likely to add value to an end to end analytic process. However, if you're a company that is attempting to predict behavior of their customers frustrated with their web experience, then they provide an excellent viewpoint into web user behavior. Web logs are giant text files that must be parsed. Each individual web interaction may be broken down as one or more lines of data within the log, and includes lots of extra records for activities that have nothing to do with user interactions. Many thousands of user visits may be included in each web log, and activities over time may be split up into many different logs. Most of the data is not useful (today, but maybe tomorrow).

Ideal location for this activity → Hadoop. For this step, Hadoop is likely the most appropriate data management platform, because of the volume and ambiguousness of the data, the web logs may be collected and persisted – and the analysis that is occurring today may produce interesting information today. Since Hadoop is an economic data store, there is less motivation to throw data away to make space for the next batch, so future analysis may likely over turn new value in the other data elements ignored today, and solve new business problems.

Step 2: Analysis

For our behavior analysis, we need to enrich that data, derive useful information about user behavior, including sessionization – this will help us monitor use patterns within web visits, and detect potentially frustrated behaviors. We also want to be able to tie this data back to our actual customers, or paying users – in doing so, we need to find appropriate customer context from other online data stores, and join those data elements together. We also want to be able to analyze which of the usage patterns represents a potentially negative usage behavior, which may require joining data over to transaction systems, or contact management sites, to see if negative behaviors were potentially reflected in one of those sites.

Ideal location for this type of activity → Analytic Data Mart for Discovery Analysis. In this case, the company will likely want to start parsing critical data elements out of the web log stream, and identify which of the web interactions actually correspond to certain customer behaviors. The web logs by themselves have no awareness of the notion of a "Web Session", so by assigning session characteristics to that behavior, the discovery platform data mart can derive patterns out of the usage behaviors – and start exposing those key data elements out into reporting accessible data objects. Of course, in order for this to be helpful, each of those user elements may have to tie back to a customer, as the rest of your operational systems are familiar with, so pulling in data from the EDW with customer elements will help that process.

This is an ideal step in the process for tools like SAS High Performance Analytics, SAS High Performance Text Analytics, SAS Visual Analytics, all would be ideal to help this part of the process.

Step 3: Operationalizing

We need to feed this data back into our operational systems, score the customers/users that are likely at risk of churn. This needs to be accessible to large numbers of customer support reps, so that they can react at real-time, when a phone call comes in, to provide potential mitigating offer. Or be factored into a specialized marketing campaign that will offer the customer something that should mitigate any potential negative action. Alternatively, we may want to include contact management logs so that future customer support reps can see that recent negative activity has occurred, and potentially route callers to specific "High Risk" support contacts.

Ideal location for this activity → Enterprise Data Warehouse. For this step, we're looking to enrich the data that is already being used for operational system support. For many customers, marketing analysis is managed in an EDW, simply because of the sheer volume of prospective customers, or by the sheer volume of marketing touches. In these case, enriching that EDW data with Churn Propensity scores from the earlier case will enable many different interface activities.

For customer contact dashboards, SAS Visual Analytics may be provide an ideal way of presenting customer data for call center representatives. In this case, SAS VA may be able to derive information from all of the different data sources, the customer information and sales transaction records from the EDW, the churn propensity scores and characteristics from the discovery data mart, and potential detail session information for supporting web interaction logs.

For marketing solutions with SAS, users may now leverage the new churn information to derive campaigns, potentially making a specialized offer to that user.

CONCLUSION

For Big Data analytics, one size usually does not fit all. The new SAS High Performance Analytics products can provide an optimized high performance model building environment to greatly accelerate large scale analytics. Tools like SAS Visual Analytics also provides an excellent platform for accelerated discovery and exploration. However, depending on the nature of the data being analyzed, the ideal platform for storage may vary significantly.

- For Big Data that is file oriented, flat, or unstructured, ideal storage may be to utilize a large Hadoop cluster node distributing the data across several nodes in a cluster.
- For Big Data that is semi-structured, like huge flat files or loosely connected together, ideal platform may be to utilize an discovery platform analytic data mart, that can be used to enrich and transform the data to derive higher value data elements.
- For Big Data that is highly structured, highly governed, used for operational decision making by hundreds across the organization, the ideal platform maybe an enterprise data warehouse, structured in an scalable MPP platform.

Or, even more likely, for Big Data that combines data from more than one of the classes above, requires both unstructured characteristics of a Hadoop platform, mixed together with the with the highly governed data managed in an EDW, the ideal environment is one that can be configured to include all of the platforms above, integrated together in an optimal managed network environment – such as the Teradata Unified Data Architecture.

SAS High Performance Analytics and SAS Visual Analytics, running on top of Teradata UDA get high performance access to the data from multiple data management platforms at the same time. Using the Teradata Fabric Based Computing model, multiple Teradata environments are attached together, onto a common high bandwidth networking environment, enabling each of the above platforms to be inter-connected, provide high speed query management, data moving, and support for both SAS HPA and SAS VA, directly on the Teradata Appliance for High Performance Analytics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Cunningham
Teradata Corporation
Danville, CA 94526

925 552 7124
john.cunningham@teradata.com