Integrated Text Analytics Teradata Environment

Chicago Technologies Incorporated

February 24, 2015

Abstract

This document outlines an integrated Teradata and SemantiGrid^{TM} environment as a platform for document storage and text analytics. Teradata is the leading Enterprise Data Warehouse (EDW) environment and SemantiGrid is the highest performing text analytics platform for semantic analysis. Combining the two environments creates an enterprise class platform for data storage and text analysis; one providing not only database functionality but also text analytics performance levels that are not possible via stand-alone SQL.

Introduction

The changing data landscape and business requirements bring new demands on the platforms and software systems that comprise our data infrastructures. Highly structured data will always be a part of the landscape and figure prominently in the data center. There are however needs for business to consider the vast amount of new data being created. Much of this is text, some of it possessing a semblance of structure (e.g., English sentences, form data), while other types being without discernible form (such as machine logs). In many instances, there is a need for management of (all) types of textual data. While it is true that it is possible to store such data within a SQL environment and that this data might be queried/analyzed using SQL, this is not generally practical especially when large amounts of data need to be stored/analyzed and when the analysis is more that just an indexed search. There are more appropriate approaches.

The BigData landscape has yielded platforms and solutions that seek to address the text analytics needs within business that fall outside of the role established by the EDW. These have solved specific needs, however, a chasm has evolved between the EDW and BigData. Even within vendor solutions where attempts have been made to integrate EDW SQL with structured environments, these have met with only minimal success. It would be instructive to understand why this is the case.

SQL Performance and Text Analytics

SQL has been designed to provide lookup of information within a tabular structure. It generally does an excellent job of this. Tables have indexes. The operating system supports performance via memory caches. Databases themselves maintain their own caches. Parallel databases extend the ability to manage very large workloads via parallel operations. So what is the issue?

While a database (especially a parallel database) performs well on lookups and joins of tabular data, the needs of text analytics differ from traditional database operations. So what is needed for text? For simple indexed lookups, a SQL approach can work well. Consider however a situation where there is a need to disambiguate text. One simple example of this might be to classify a document and determine wether a particular string might be contained within a document store. This is where the system must consider more than just an indexed lookup. Perhaps we need to consider synonyms within both the source document and the search string. This easily equates to an M*N*R where M is the document word count, N is the number of words within the search string and R is the average synonym count.

While simple and synonym-augmented lookups can make searches over a large document store be prolonged and compute expensive, more advanced and accurate searches can increase the computational requirements exponentially to where a solution is no longer viable. Consider a situation where simple word search is simply not sufficient to meet a document search. One situation might be the need to consider bag of words such as an entire sentence to establish context. In this case, there is a need to ascertain the meaning of words based on their inclusion in the sentence.

Consider something like: “The slugger stepped up to the plate and ripped the pitch into the right field stands much to the dismay of the entire Tiger bench.”. There are many challenges with disambiguation of this sentence. There are so many different meanings of “slugger, plate, ripped, pitch, right, field, stands, Tiger, bench”. To perform an accurate match, there is a need to consider inter-relationship between the words and establish their true meanings. Once this is known, we can apply the same principle to the target documents. The scope of this type of (more accurate) text analytics/search is profoundly more compute intensive. Attempting to solve this with SQL can be an extremely difficult and expensive proposition.

Regardless of indexes (these will help), there will be a large number of SQL operations required to satisfy the large number of search operations. Each SQL operation has a performance footprint that is encumbered by many operating system process switches, system calls and database processing overhead. Add to this the potential need to access disk and a SQL operation typically takes milliseconds to complete. As the data store grows and there is a continued need for more precision via semantics, the overhead of depending on SQL, becomes an expensive proposition both in terms of performance and system utilization/cost.

The SemantiGrid^{TM} Approach

SemantiGrid^{TM} has been designed to address the core issue of semantic Natural Language Processing (NLP); word search and semantic processing. From a platform and solution perspective, SemantiGrid^{TM} forms a complete solution with features such as: massive storage capability (beyond Hadoop), HIPAA and PCI compliant security, fault tolerant operation (self healing), multi-site capability for 0 data risk, automated operations for all NLP preprocessing functions. From a performance perspective, its foundation is an in-memory multi-dimensional knowledge base.

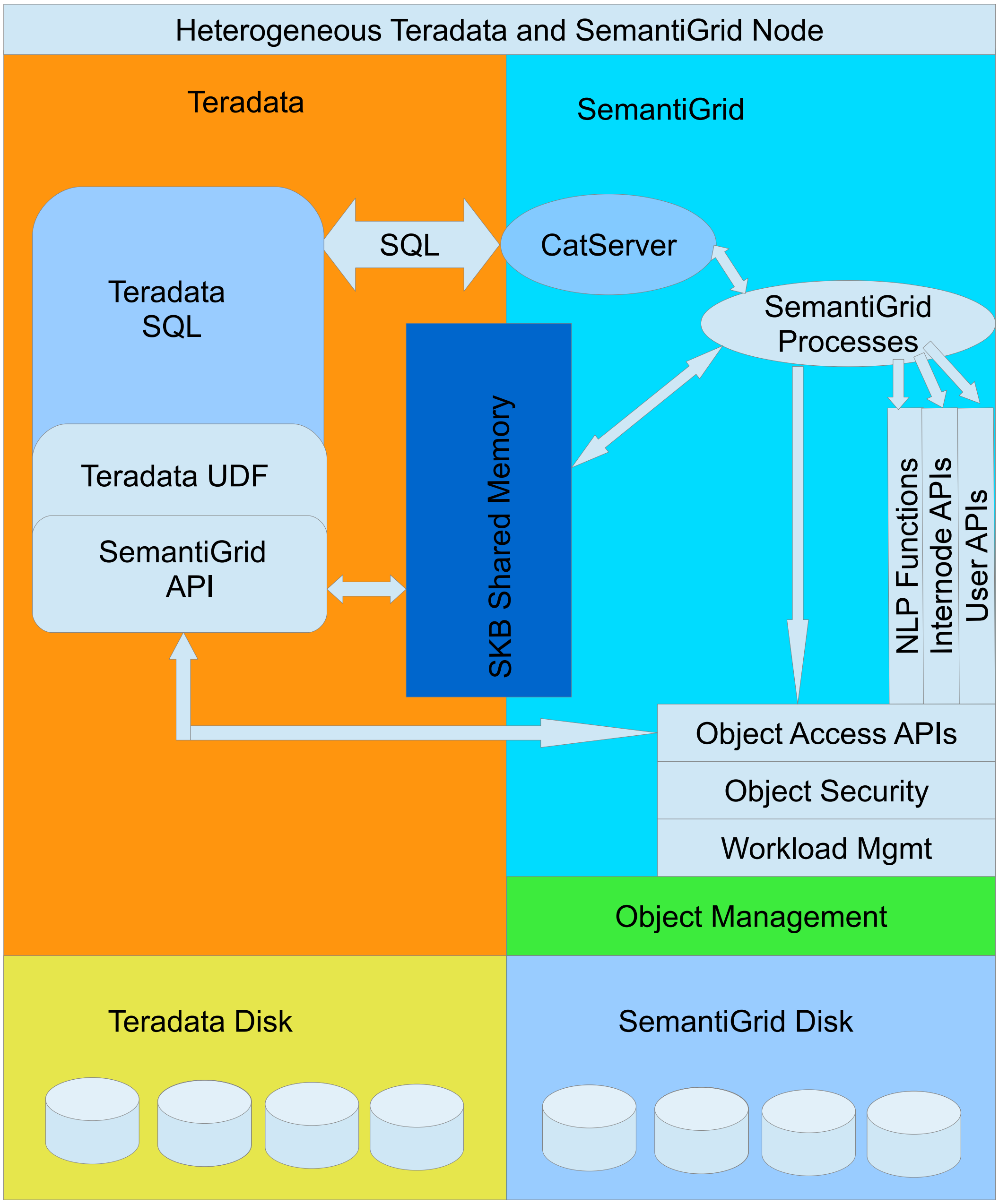
The SemantiGrid Knowledge Base or (SKB^{TM}) is a foundational technology for any serious semantic processing endeavor. While in-memory processing is at SKB^{TM}'s core, the semantic functions and the underlying SKB^{TM} implementation allow it to provide performance that is at least 3 to 4 orders of magnitude higher than any SQL database (including parallel databases) for equivalent operations. The SKB^{TM} can achieve very high performance due to a number of features including the N-Dimensional knowledge structure and efficient access mechanisms to the knowledge data. Having the entire knowledge base in memory is another consideration but so is the ability that allows (many) tasks to share the same copy of the knowledge base. Ultra performance is achieved because SKB^{TM} functions each complete their operation without the need to give up control to the system, thereby eliminating significant latencies and providing a highly efficient implementation.

Teradata and SemantiGrid^{TM} – An integrated Platform

In any large scale document database, there is a need for traditional database functions. This is because there are many operations on documents that require access to meta data about the documents stored within the document database. There is also the potential to run analytics on information that might be stored within SQL tables that has been preprocessed by the SemantiGrid^{TM} platform. Analysis can in fact benefit from both SQL and the information stored with SemantiGrid^{TM} and the SKB^{TM}. This can be accomplished using SQL with the help of UDFs that connect directly to the SKB^{TM} and via APIs to SemantiGrid^{TM} functions. This also permits the use of modern BI tools and interfaces to utilize advanced text analytic capabilities to yield highly performant and functional solutions.

The basic approach is to store documents within SemantiGrid^{TM}, have SemantiGrid^{TM} preprocess documents, extracting information and building an information store on each document. Most of the preprocessed data will remain within SemantiGrid^{TM}, however, meta-data about the document will be stored within the Teradata Database. This will provide very fast access to information about documents along with the ability to perform semantic searches - all within SQL. There are however UDFs available for processing to extend the capabilities of pure SQL. This processing can be much more efficiently handled by SemantiGrid™ especially due to the SKB™ and because SemantiGrid™ has preprocessed document data using a variety of textual methods. The end result is the ability to store massive amounts of document data and the ability to efficiently search, query and analyze document text.

Consider the below diagram:



The diagram above shows a tight integration between Teradata and SemantiGrid™. The key is to provide access by Teradata UDFs directly to the SKB™ and a very high level of performance. This is accomplished by Teradata and SemantiGrid™ sharing a Teradata node. There is no requirement however that SemantiGrid™ share nodes and it is also possible to have SemantiGrid™ and Teradata run on separate platforms. It is also possible to use a subset of Teradata nodes as hosts for SemantiGrid™.

The end result is a Teradata platform that becomes a very capable text processing system.

Applications – The Key To Success

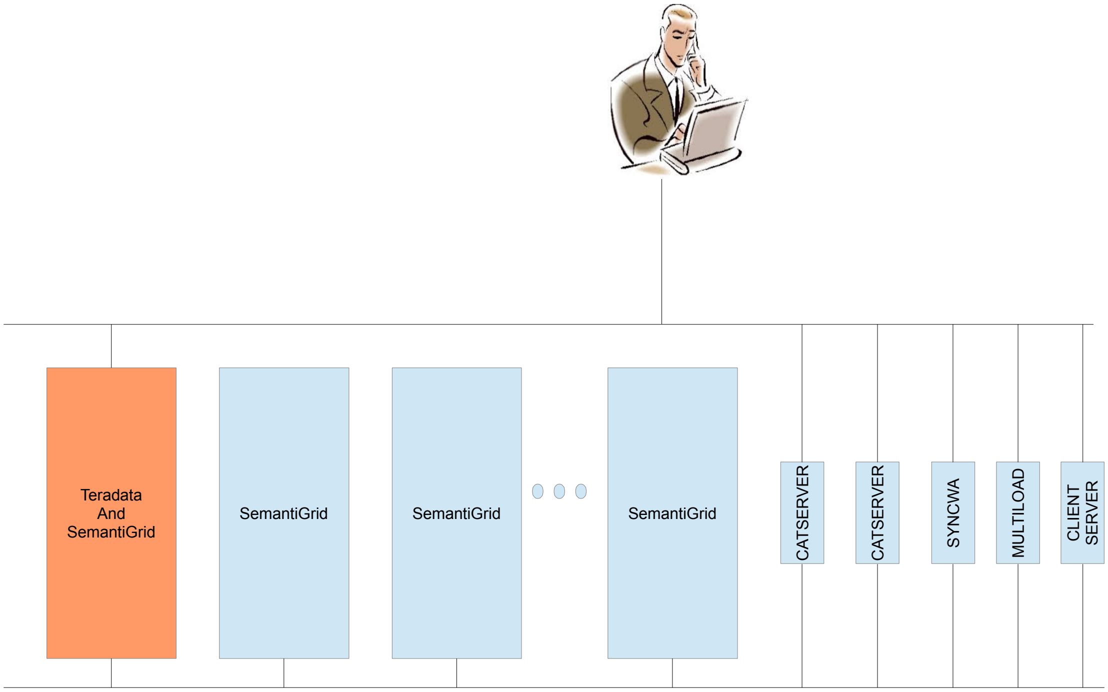
A platform such as can be achieved with the integration of Teradata and SemantiGrid™ is an attractive combination to many potential customers, however, there are additional opportunities beyond the platform itself. The explosion of data (90% of all data ever generated has been generated within the last two years), brings a need to store massive amounts of and then to have the capability to efficiently query it. There is no system currently on the market that meets the requirements for this level of data retention and access – SemantiGrid™ provides this capability.

Let's look at just one aspect of this by considering the mandate that all court systems have of storing court cases. Most documents sit in warehouses in boxes on pallets. There have been attempts to scan and place these in document management systems. These have largely failed (Cook County of Illinois being one of these). The challenge with any of these implementations is that not only must the data be stored and protected, but the system must be efficiently accessed/queried (forgetting analytics for a moment). This is a major opportunity, however, there are many others.

The same challenge to store/retain data exists within the corporate world and within government. Many of these are similarly an opportunity to store large volumes of data. Consider the healthcare area where there is a need to store massive amounts of data. Much of this is text, such as doctors notes and patient information. While attempts at using platforms such as Hadoop, MongoDB and Elastic Search have been made, none of these meet the requirements for Scaleability, Performance, Availability, Reliability, Cost-effectiveness, and Security (SPARCS). A solution utilizing a heterogeneous approach with Teradata and SemantiGrid™ does meet and exceed these requirements. The need by government to store, access and analyze large amounts of data are yet another situation requiring the same SPARCS features.

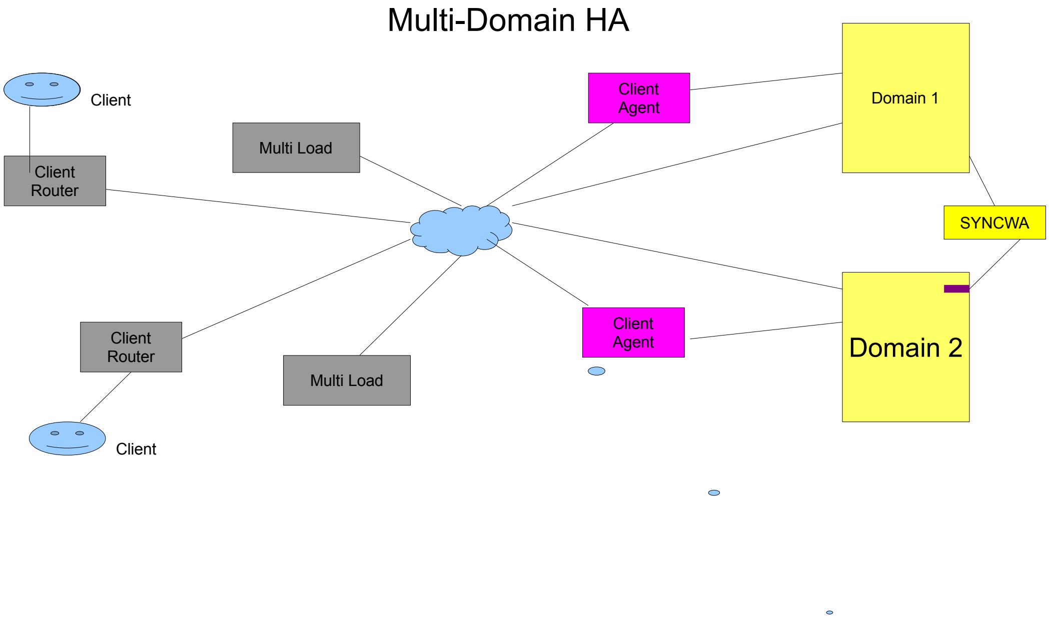
A SemantiGrid™/Teradata environment is not just for BigData. It can provide a very cost-effective solution for a corporation needing to store business documents, email and server logs. These can be cost-effective because the majority of the hardware can be SemantiGrid™ low cost nodes and storage with a smaller Teradata footprint. There is no requirement that there be a 1-for-1 coexistence between a Teradata node and SemantiGrid™ node. The below diagram shows a possible configuration where a smaller Teradata system is mated to a very large SemantiGrid™. Please note that a heterogeneous Teradata/ SemantiGrid™ node provides the high level of performance that is expected from a collocated SKB™ and Teradata node(s).

Additional components shown in the below include: CATSERVERS, SYNCWA and CLIENTSERVER. The CATSERVERS are replicated catalogs that hold the operational metadata for the system. A CATSERVER can be resident on the main Teradata processing nodes or can be (as displayed) on standalone SemantiGrid™ nodes that contain a single node TD database or Teradata Appliance. The CATSERVERS are redundant but SemantiGrid™ has been designed to function without a functioning CATSERVER. Lack of a CATSERVER will impact the performance of the system but SemantiGrid™ can communicate directly with its nodes to fulfill requests normally handled by the CATSERVERS until they are back in operation.



High Availability and Fault Tolerance

SemantiGrid™ has been designed to be fault tolerant. Fault tolerance is not only at the cluster level but extends to geographically distributed domains. SemantiGrid™ has been designed to provide multi-domain availability and synchronization. The below diagram shows two components that comprise



synchronization capability of SemantiGrid^{TM}. First the Client Router. This has the ability to provide fault tolerant access to multiple domains so that users are not subjected to service interruptions due to a domain or its network path being unavailable. There are also “Multi-Load” components that allow parallel loads/updates to multiple domains. An object creation/upload can be directed to as many domains as are configured. This provides true 0-data risk operation without the need for a secondary data backup capability outside of the multiple domains. (Note: Each domain is in itself a fully fault tolerant cluster).

Each domain however is autonomous in operation and the external services are responsible for provisioning data to each of the domains. Inter-domain replication is a potential need so there is an inter-domain synchronization capability (syncwa) that can be tasked with ensuring that objects are consistent across domains.

Database Consistency

The SemantiGrid^{TM} consistency between domains (for SemantiGrid^{TM} data) is complemented with a database level consistency to ensure that two tables within different domains or within two different Catalog Servers (CATSERVERS) contain the same information. This is primarily for control information but can be extended for other tables as well. This is a patented fast check and re-sync capability that permits very low latency and low overhead table consistency checking for the database.

Summary

SemantiGrid^{TM} brings an application capability to Teradata that allows it to address an entire market for large data storage and analytics. This market is just coming into fruition. The timing is right. There is much interest and competition for this type of solution. The combination of the finest and most performant SQL database in the world and the fastest and most functional text analytics platform in the world are a winning combination whose time has come.