Running Head: APPLIED KNOWLEDGE MANAGEMENT

MetaTech Consulting, Inc.

White Paper

Application of Knowledge Management Constructs to the Massive Data Problem

Jim Thomas

July 23, 2004

Application of Knowledge Management Constructs to the Massive Data Problem

As the scale of centrally managed data stores increase from gigabytes (10^9 bytes of data) to terabytes (10^{12} bytes of data) to petabytes (10^{15} bytes of data) and beyond, the challenge of accessing and managing the stored content increases as well. Historically, the common practice used to deal with this phenomenon was to increase the systems administration staff in equal proportion to the increase of data being managed. Such practice is evident in the resource estimation technique commonly employed by Information Technology Managers that derived required manpower by applying a function to the anticipated capacity of the system. For example, at the end of the last decade it was reasonable to estimate three administrators would be required for each terabyte of data.

For previous decades, data management technologies matured roughly on par with the growing storage demands such that as capacity increased it was possible to maintain a constant number of administrators. This encouraged organizations to continue the practice of storing all data indefinitely.

During the past several years, it has become apparent that the rate of increase of required storage capacity is now significantly outpacing the rate of maturity for the administration tools and technologies. The result is that the number of administrators is now increasing at rate proportional to the demand for storage capacity – estimated to be approximately an order of magnitude each 18 months. Maintaining such a steep increase of manpower has proven to be impractical.

Background

Massive Data is the term applied to collections of data of great volume – generally measured in terms of at least *PetaBytes* (i.e. 10^{15} bytes). While simply storing and accessing such quantities of data provide a daunting technical challenges, harvesting the tacit value from these repositories is completely overwhelming.

There are three alternatives that can offset this trend: a) some dramatic increase in capability of information management systems through technology advancement, b) halt the trend of producing more data, and c) become more judicious in deciding how long to persist (e.g. store and manage) content. It has been maintained by this author that, irrespective of the first two alternatives, the third alternative deserves serious consideration. This assertion is based on the thesis that while some data is of inherent value, other data is not and that most, if not all, data diminishes in value over time.

The state-of-the-art of information management systems is such that it is neither possible to ascertain the value of content with some automated of semi-automated process nor to manage content based on *value* even it were known. Therefore, there exists a need to undertake a rigorous study of the related factors and technologies to synthesize some solution with these capabilities.

The scientific and engineering communities exercise applications that execute on sets of data for the purpose of discerning answers to our most challenging problems. The datasets are frequently formed over extended periods of time (e.g. historical records), from sensors that respond with incredible frequency (e.g. real-time flight instrumentation or integrated sensor networks), from sensors that have tremendously broad bandwidth (e.g. spacecraft), or a combination of all three. As the magnitude of the problem has

shifted from megabytes to petabytes, it has been termed massive data. Further technological maturity is critical to our efforts at exploiting, to the greatest extent possible, the information, knowledge, and intelligence hidden within these massive data sets.

Data, information, and knowledge are often critical to successful business operations. To varying degrees, and at varying stages in their processes, businesses both produce and consume these artifacts. Information technology systems of a business process both constrain and define its capacity to exploit these artifacts over time. Often, even with state of the art technology and expansive budgets, businesses find themselves unable to effectively store, manage, and process all available data – together with information and knowledge – and either actively or passively they must allow the excess to perish.

Each instantiation within each business domain has differing capacity to deal with data which ranges from gigabytes (10^{9} bytes of data) to terabytes (10^{12} bytes of data) to petabytes (10^{15} bytes of data). It is this largest scale of data problem that is generally acknowledged as massive data.

Architectural Considerations

Any viable solution to the massive data problem must address a number of factors concurrently. Though not an exhaustive listing, a meaningful set of factors is presented presently to provide the audience with a context for subsequent discussions. They include: a) content value, b) risk of discarding content, c) and characteristics of storage media types. Following a brief treatment of each of these, the architecture itself is described.

Content Value

The assertion: Over time, content diminishes in value. An example from the telecommunications industry is provided for illustrative purposes. When a telephone call is placed, a call records is created. This record will be used to calculate billing charges, for dispute resolution, and as evidence in court if justification warrants. It has significant value to the organization for some period of time. Over time, however, the individual call record becomes less valuable to the organization. Once the bill has been issued, the window for filing a dispute has passed, and the legal retention period has come and gone, it is difficult to argue the call record has *any* value remaining.

It is certain that some statistical information derived from the collection of call records – of which the record is a part – will continue to have value. The company is certain to have interest in maintaining knowledge of how usage changes over time. As an example of the value of summarized or aggregated data, one need only consider how the telecommunications industry is assessing the impact of cell phones on the traditional telephone market.

It is not suggested that the value diminishes at a regular or linear rate. In many cases, such as in the aforementioned telecommunications example, the value diminishes at some irregular interval in the form of a step function. A useful characterization of an efficient Information Management System is one that has the greatest ratio of valuable content to total content.

Risk

In this context, the term *risk* is used to mean the probability that there will be a need for content after it is permanently deleted from an information system. Intuition

suggests that the risk associated with expunging content varies with time in proportion to the value of given content. This implies that, in general, the risk of purging data is greatest immediately after it is received and that it decreases with time. A second useful characterization of an efficient Information Management System is one that has the greatest ratio between content with a high probability of use to total content.

Media Capacity, Cost, and Performance

When architecting an Information Management Solution, one must consider the media onto which the content will be stored. In solutions designed to scale to support massive data, solid-state (e.g. memory or RAM), hard-disk drives, and magnetic tape cartridges will each likely be used in some proportion. Though solid-state media has the most desirable performance characteristics, it would be prohibitively expensive to build even a multi-terabyte storage solution only this type of storage media. Disk-drives have been the compromise between solid-state and magnetic tape cartridges which are less expensive but lack the performance requirements of online processing systems.

The status quo of the storage hierarchy has become less rigid in recent years with the advent of robotic tape libraries that stripe data across multiple tapes cartridges to achieve an aggregate performance approaching that of disk-drives. Techniques like this one have stratified both the disk and tape markets into two layers: high performance and low performance. An additional characterization of an efficient Information Management System that is useful is one that has the greatest ratio of the measure of performance and the least total cost of media. A measure of performance must include satisfaction of user’s requests, ingest rate, access rate, reliability, availability, and sustained transfer rate.

Proposed Solution

The remainder of this paper presents a notional Information Management Architecture (IMA) as illustrated in Figure 1. A specific treatment is given for certain key components of the architecture together with a discussion of the *Knowledge Management* principles that are considered to hold promise in mitigating unwieldy increase of capacity.

Data Cleansing and Preparation Function

Data arriving for ingest into an IMA is frequently of inconsistent quality. Furthermore, data often arrives without regard to sequence dependencies that might exist independent of any given information feed. For example, information may arrive as streaming data from an online transaction processing system with information on purchases at point of sale terminals. A second feed into the system containing account summary information may arrive as the result of a batch process run at the end of each business. If a customer established an account on the same day as they performed their initial transaction the transaction would arrive at the IMA prior to knowledge of an account.

For this reason, and countless others, data presented for ingest into the IMA is staged in a *cache* until all requirements for ingest are satisfied. As the *Cleansing and Preparation* function can be CPU intensive, the *cache* also functions as a buffer to prevent data loss during periods when it is arriving quicker than it can be ingested. *Cleansing and Preparation* is driven by a set of business rules that, though complex, are well understood and tractable.

Classification Function

In an attempt to minimize the total capacity of the IMA and to ensure optimum usage of the storage media used, the IMA classifies content during ingest. The *Classification* function assesses given characteristics (e.g. source, quality, uniqueness) of the arriving data against rules to determine the best period of retention and media. The content is tagged with this information such that the *Cross Media Manager* can store the content appropriately. The aging rules result in a *purge-on-date* that can key the storage system when it should perform its *garbage collection* to reclaim capacity. Similarly, the migration rules result in a *media identifier* and a *migrate-date* that is used by the *Cross Media Manager*.

Aging Engine

As the rules that drive the decision of when to age data from the system may change over time, the storage systems engages the *Aging Engine* to reevaluate the *purge-on-date* prior to expunging the content from the system. If the new rules dictate, the *purge-on-date* is adjusted to some later date. Otherwise, the *Aging Engine* provides notice to the storage system to include the content in the next *garbage collection* activity. Metrics are forwarded from the *Aging Engine* to the *Rules Management* function for evaluation.

Migration Engine

Similarly, the rules that drive the decision of which media to use to store the data may also change over time. The *Cross Media Manager*, the storage system component responsible for placing and locating content within the storage system, engages the

Migration Engine to reevaluate the *media identifier* and a *migrate-date* prior to moving content between storage media. If the new rules dictate, the *migrate-date* is adjusted to some later date. Otherwise, the *Migration Engine* provides notice to the *Cross Media Manager* of the new *media identifier* and a *migrate-date* such that it can move the content to the target media. Metrics are forwarded from the *Migration Engine* to the *Rules Management* function for evaluation.

Cross Media Manager

As previously mentioned, this is the storage system component responsible for placing and locating content within the storage system. Additionally, this component provides metrics on the capacity and composition of the total and available storage system to the *Rule Management* function. This information will influence the aggressiveness of the ageing and migration rules.

Access and Query Management Function

This function facilitates access to stored data by users (e.g. people as well as processes). This component provides metrics to the *Rule Management* function on the satisfaction of queries as well as on the queries themselves. This profile information will also influence the aggressiveness of the ageing and migration rules.

Rules Management Function

This function aggregates the metrics collected from each of the other functions within the IMA for the purpose of optimizing the aging and migration rules. Unlike the other functions that are driven by well understood and tractable rules, the *Rules Management* function does not. In traditional information management systems, this

function relies on heuristics and the intuition of the systems administrator. In the IMA, a more sophisticated technique to rules management is employed. This techniques, together with that used buy the other functions, are treated presently.

Application of Knowledge Management Constructs

This section of the paper partitions the previously discussed components of the IMA based on meaningful characteristics. The first set of components is characterized by the ability to derive a result through well understood rules. The remaining set of components is characterized by the lack of tractable rules to derive a required result. Each set is discussed as an application of the Knowledge Management construct that best satisfies its demand. The two constructs that will be covered are *Expert System* and *Neural Networks*. The discussion will conclude by addressing the processing modalities issues relevant to the sets of data and the appropriate constructs.

Application of the Expert System Construct

Most of the functions of the IMA can be driven by well understood and tractable rules. These include a) the Data Cleansing and Preparation function, b) the Classification function, c) the Aging Engine, and d) the Migration Engine. For each of these functions, it is possible to codify the business rules and to execute them in some optimal way. Though these rules may be complex and may, at times, appear counter intuitive, they are solvable. This common characteristic of these functions support the architectural design decision to solve them through *Expert System* technologies and techniques. Specific examples of the use of Expert Systems approach include the assignment of data to a given class within the *Classification* function and determining the optimum retention period assigned to each class of data.

Application of the Neural Network Construct

The *Rules Management* function, having historical data to evaluate rather than tractable rules, is not an ideal candidate for an *Expert Systems* solution. Rather, a *Neural Network* solution is an ideal implementation for the challenge of optimization of the rules. Each set of data ingested into the IMA results in a set of metrics. Likewise, on each occurrence of data being purged (or consider for purging) or migrated (or considered for migration) metrics are produced. Also, with each query by users, further metrics are generated. These processes result in a large set of the historical. The *neural network* solution can mine through the metric to determine if the rules for classification are correct and if the rules for aging and migration for each class of data are optimum.

Processing Modalities

As the processing is tremendous for complex *neural networks* on deep sets of data, the IMA *Rules Management* function would be conducted offline rather than in real time. The *Expert Systems* rules, on the other hand, can be efficiently evaluated at online processing rates. Where the ingest rates are extreme, the rules can be applied in a massively paralleled scheme. Furthermore, the rules can be encoded in hardware such as with *Field Programmable Gate Arrays*. Though this solution is costly, it is feasible in some of the most demanding large scale implementations.

Conclusion

Information Management Systems designed to solve the needs of Massive Data require a novel approach to architecture. It is necessary to construct a set of functions that employ an optimized set of rules that provide an automated or semi-automated management scheme minimizing the quantity of unneeded content persisted and the most

appropriate utilization of the media used. With an understanding of the functions involved, it was determined that an *Expert Systems* approach was the most suitable for executing the various rules and that a *Neural Network* approach was the appropriate solution for ensuring the rules remained optimum over time.

Figure 1

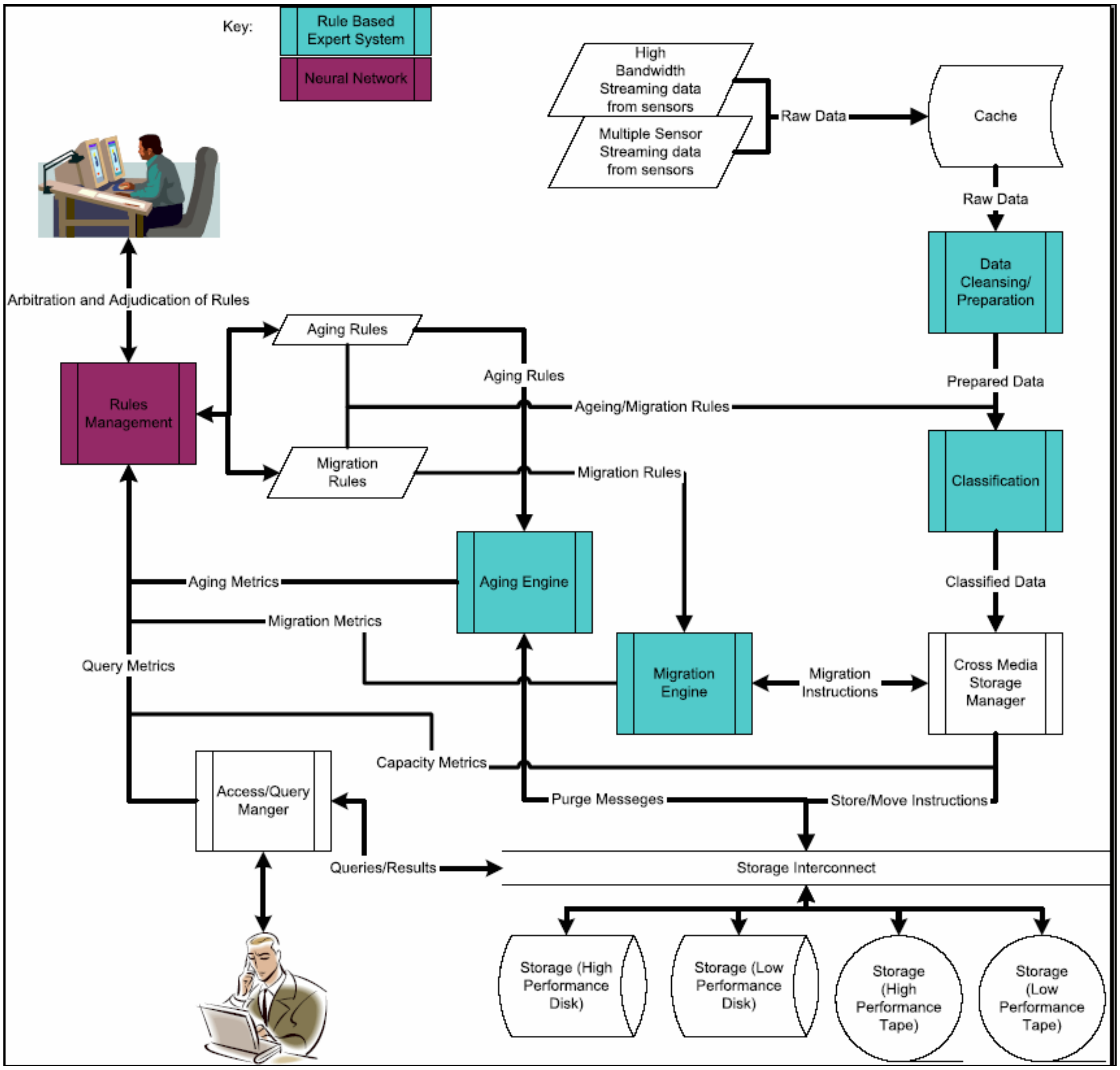

Figure 1. Notional Profile of Workforce Experience