Running Head: ADDRESSING THE MASSIVE DATA PROBLEM

MetaTech Consulting

White Paper

A Characterization of the Massive Data Problem with a Discussion of the Suitability of the Data Warehouse as a Solution Component

Jim Thomas

November 26, 2003

A Characterization of the Massive Data Problem with a Discussion of the Suitability of the Data Warehouse as a Solution Component

The purpose of this paper is to provide an overview of the *massive data problem* to illuminate the need for means of managing vast amounts of data for the benefit of some enterprise. The concepts and technologies of data warehousing will be provided along with the suggestion that they have some degree of suitability as a component of a viable solution to the overall massive data problem. The pros and cons of the identified technologies are revealed and discussed briefly. The paper concludes with a proposed approach to addressing the massive data problem with three specific areas where additional research and development is needed.

Characterization of the Massive Data Problem

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 supporting business processes 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 content to perish without fully exploiting it.

Each instantiation within each business domain has differing capacity to deal with data. Small and medium sized challenges range in size from *gigabytes* (10^9 bytes of data) to *terabytes* (10^{12} bytes of data) while large systems have capacity demands of *petabytes* (10^{15} bytes of data). Carino, Kaufmann, and Kostamaa (2000) state that solutions are now emerging that may well extend into the *yottabyte* (10^{24} bytes of data) capacity range. The term *massive data* is generally reserved for the petabyte and larger problems.

Massive data is produced from three classes of sources: a) deep archives, b) wideband data collectors such as those employed in space borne sensors, and c) integrated networks of disparate sensors. Each of these is addressed in turn though, with consideration to required brevity, a thorough treatment will be omitted. Interested readers are referred to the cited references for a more complete appreciation of the material. In the interest of brevity, the treatment of the three classifications of massive data sources is deferred for later work.

Characterization of Data Warehousing as a Solution Component

The coining of the term *Data Warehouse* is generally attributed to Mr. W.H. Inmon. Inmon (1996, p.3) defines the data warehouse as "a subject oriented, integrated, non-volatile collection of data in support of management's decisions." The principles of data warehousing provide the foundation of numerous commercial products available for application to a growing set of challenges. While initially envisioned as having principal applicability to the *Decision Support Systems* (DSS) problem set, it is now present in many *Online Transaction Processing* (OLTP) and *Online Analytical Processing* (OLAP) systems.

Rather than being a hard technology – one that is discrete and unique from other technologies – data warehousing is more of an innovative approach to the application of existing technologies. Physically, a warehouse consists of processing and storage components. Content (e.g. data, information, knowledge, etc.) is persisted on storage media within a defined *schema* that is optimized to support the business needs. An operating system, database environment (generally a relational construct), applications, and tools are the principle software components that perform the warehousing functions.

Content is prepared for ingest into the warehouse through a process known as *Extraction, Transformation, and Loading* (ETL). Though the ETL process, erroneous data is corrected, missing data is resolved, and duplicate data is reconciled to produce the *data of record* for historical reference. Additional processing may include *aggregation* or *summarization* of the data is statistical details are valued by the business at hand over the discrete data itself.

Once ingested into the warehouse, the content in the form of atomic data (e.g. discrete data) or aggregated information is available to the business processes for exploitation. As previously mentioned, the information is persisted in the database in a schema optimized to the business concerns at hand. While the data elements themselves may be predetermined and will limit the possibility of what the warehouse can provide, the business needs that of the enterprise will dictate how the schema is optimized. Figure 1 illustrates the diverse access and retrieval demands that may be placed in a warehouse. Furthermore, the warehouse must support information retrieval on demand, with tolerable performance constraints. To accomplish this, the schema will often contain fields to hold information computed or derived from the source data during the ETL process. Further analysis of data warehousing if provided in the following paragraphs.

Analysis of Technology

Pros

The data warehouse has matured as a robust set of capabilities functioning within a framework that enables the orderly persistence of large quantities of data and information. The tools and services of the warehouse provide efficient access to and manipulation of the data and allow the users to perform tasks ranging from operational and decision support to mining and exploration. As data warehousing technologies become more capable, and as the business analysts, administrators, and executives become familiar with them, new and innovative techniques evolve to better leverage existing information content (e.g. data, facts, etc.) to the benefit of the corporate bottom line. Warehouse-related technologies support *Business Intelligence, Customer Relationship Management,* and other practices that are employed to protect and improve market share. Common to these routines is a need for an integrated and deep view of a business' data to determine trends, relationships, and patterns.

Improvements in warehousing technology have continued in the areas of a) hardware, b) operating systems, c) database engines, and d) applications. The following paragraphs provide a brief discussion of improvements that have been made in each of these areas. The focus of the discussion will be on the impact the improvements have on warehousing in particular.

Hardware. Though many warehouse environments have been successfully deployed on general purpose hardware platforms, those demanding the greatest performance and scalability pose a challenge to hardware vendors. Servers, storage, and interconnect bus structures together provide the infrastructure on which a warehouse solution is built. *Teradata* ™, a leader in the warehousing solutions market, provides a worthy example of a vendor that considered each of these hardware components to improve their warehouse offering. Balliger (2003) asserts the ability of the *TeraData* ™ platform is capable of achieving linear scalability as data volumes grow (Figure 2) by employing a proprietary, purpose-built communications interconnect supporting the servers and storage.

Operating systems. The demand that warehousing applications (e.g. databases engines, analytical tools, etc.) place on computing systems is extreme. Data sets that are many gigabyte in size must be loaded from disc or tape into memory for manipulation. Sorting algorithms executing on these large data sets is a single example of a challenge for the operating system supporting the warehouse. Operating systems of the past were incapable of addressing the terrific volumes of memory or the number of physical storage devices commonly used in larger warehouses.

Database engines. Relational, object, and dimensional database engines have continued to mature in path parallel to warehousing. While not sufficient to accomplish warehousing, a database engine is absolutely necessary. The architecture and design of the database system used within a warehouse imposes constraints on its performance and scalability. Likewise, the intended application of the warehouse drives the selection of the database. For example, a warehouse intended to function principally as the engine of a decision support system will likely be built around a multidimensional database. These databases support complex data structures optimized for well defined queries in a very timely manner. This is contrasted to the relational database that, when well designed, provide rapid response to ad hoc queries. Vendors have tailored their products to better suit the demands of large warehouse implementations by focusing on the different and distinct information challenges being addressed by warehouses. As illustrated by Strange (2003), the principal database management system vendors have continued to mature consistently (Figure 3).

Applications. A variety of vendors now offer tools, utilities, or other software that support warehousing. A shift in the commercial software market towards open, standards-based development has resulted in a far greater level of interoperability than previously experienced. Standards for data interchange, in specific, have greatly aided warehouse solution developers. The *eXtensible Markup Language* (XML) provides a self-describing structure by which disparate applications can exchange data without needing intimate insight to how the other applications represent the data.

Cons

Though there have been significant advancements in technologies that support data warehousing, shortfalls continue to exist and to hamper efforts to address the massive data problems of tomorrow. The following paragraphs provide a concise treatment of a few of the challenge areas that are not satisfactorily resolved by the warehousing solutions of today. Those to be addressed in turn include a) unstructured data, b) heterogeneous data, and c) streaming data sources.

Unstructured data. The data and information artifacts of concern are no longer limited to well defined sets of textual (e.g. alphanumeric) characters within reports or forms or other well defined formats – *structured data*. Rather, there is a strong business case to include *unstructured data* in data warehouses. Few organizations have continued to remain competitive while dictating strict, inflexible, unchanging adherence to defined data. Data communicated through text messaging and emails – communications devices that have contributed to improved efficiency in many organizations – provide both structured and unstructured data. Message header information (e.g. that defined by the appropriate protocol) is necessarily well structured while the message payload (e.g. the message being transmitted) is generally unstructured. It is undesirable to manually extract the message content from an email (possibly sent from a customer to a service representative) to convert it into a structured form for use by a customer relations management (CRM) data warehouse. Only limited progress has been made in developing systems that are capable of performing this task automatically.

Heterogeneous data. To build on the previous CRM example, it is useful to consider an instance when the content is not only text. Sound, animation, pictures, and video can be more effective than text for communicating information in certain situations. An email sent by a customer might include a digital picture of a failed part or of damage that resulted from the failed part. Though storage systems are capable of persisting any form of digital information, they must also manage the content in such a way that it retains (or even adds) value to the business. Many warehouses of today store unstructured, heterogeneous pieces of data by storing them in some binary format. This approach is inefficient from a capacity management perspective and also lessens the value of the data. Improvements in processes for identifying, classifying, and managing heterogeneous data are needed.

Streaming data. Traditional data warehousing solutions rely on a phased approach to ingesting data. Generally, data was migrated from the operational systems to a staging area – a temporary storage area where it was held until all dependent data had arrived and the warehouse was ready to handle the new data. While there were other reasons for the staging area, these two are of particular interest as both are obstacles to the objective of ingesting data from flowing streams rather than from discrete flat files or records that arrive in discrete intervals.

The massive data problem compounds the need for handling streaming data. As stated previously in this paper, sources of massive data include those sensors with very wide apertures (e.g. space born sensors) as well as those sensors that are integrated to form large networks. In either instance, the data may arrive for ingest by the warehouse at rates and volumes that make staging the data for any period of time unreasonable – the backlog simply grows too fast. Also, as the nature of the users of the warehouse changes, batch processing of staged data becomes unreasonable. As the warehouse becomes more functional as an Online Transaction Processing (OLTP) environment (such as the reservation agents for airlines), near-real-time visibility to integrated data becomes critical to business operations. As with making backups, the windows to perform the task of ingest is becoming ever smaller. There has been only limited progress in handling streaming data and ingesting data without taking the warehouse off-line.

Proposed Approach to Addressing the Massive Data Problem

After examining the massive data problem together with the technologies that are regarded as promising contributors to the remedy, this author has identified three areas deserving focus and directed attention: a) metadata, b) ontology, and c) indirect management of the data. A concise discussion of each follows presently.

Metadata

The efficiency of an information system is bounded by the metadata that it employs. Though it is true that modest databases can satisfy user requirements for access with few challenges to the database designer or administrator, it is also true that as the database grows in capacity or complexity the efficiency of the design becomes critical to usability and maintainability. For very large databases, particularly those functioning as warehouses, improved performance is achieved through heavy use of indexing and creative use of metadata. Such is particularly true when the warehouse is used for data mining, knowledge discovery, or other exploration tasks.

Metadata has still more utility beyond improving access performance. Few data warehouses ingest data form a single operational system. Rather, warehouses generally receive data from many different operational systems. Each the source system possess its own data models and each has a level of stability and maturity that results in some level of change. Accurate metadata capturing specific source of each piece of data as well as the particular version of the source systems data model will enable the warehouse to provide data reliably as the source systems – together with their data models - mature and change over time.

Ontology

It is common for the source systems contributing to a warehouse to have some common context – the business, the product line, the customers, etc. Across the enterprise, in the minds of many users, there typically exist a number of disparate mental models that together form a representation of the enterprise as a whole. This model, known as an *ontology*, consists of some number of entities descriptions together with the description of the relationship between each. In the minds of the uses, the data from the disparate data sources make sense. The business functions only as long as the collective brain trust is functional and remains in place.

An observation of the office workers and executives of any large enterprise would suggest that no single individual possesses a complete and accurate mental model of the enterprise. Furthermore, few individuals have the intellectual capacity to comprehend the nuances of anything beyond a fairly simple model – a limitation not inflicted upon modeler computing systems.

It is suggested that by modeling the enterprise – by generating an ontology for the context of the enterprise – and codifying this for use by applications and tools functioning on the warehouse data it would be possible to harvest far more knowledge from the data that already exists. Complex relationships that exist unnoticed by so many workers that have neither the access nor vision to see that which exists before them can be determined and used for the benefit of the enterprise.

Indirect management of data

As the size of warehouses increase, they reach a point where it is not longer feasible for it to function – for example, the time it takes to generate an index following the ingest of a set of data exceeds that time that the system is permitted to be off line to perform that function.

Conventional wisdom has dictated that effort (indexing) should be shared between more processing resources to accomplish the task in the allotted time. There comes a point where even this approach fails. The massive data problem will prove this point.

It is suggested that an alternate approach to the problem is to generate an abstraction of the data that is sufficiently detailed to provide necessary functionality though it is also considerable less in volume. This abstraction of the data, a form of metadata, allows the processes to manage and access the data indirectly. Furthermore, it is asserted that if only this metadata is persisted in a database and that the source data is persisted as an artifact in a flat file management system, the processor resources necessary for accessing and managing the data would be considerably reduced. For example, a dataset produced by a photography sensor onboard a space borne may be in excess of 50 gigabyte in size. Metadata reflecting the location in the file system of the source file as well as the descriptive information of when, where, how, why the image was produced as well as what it is of could likely be stored in only a few kilobytes and could easily be managed in a database as structured data.

Conclusions

This paper has presented both an overview of the massive data problem and a treatment of the state of the art of data warehousing. The pros and cons of data warehouse technologies were addressed. It was asserted that additional development of the existing technologies, with attention to metadata, ontologies, schemes for indirect management of massive data sets, is required to adequately address the challenges of massive data.

References

Ballinger, C. (2003). *The Teradata database scalability story*. Retrieved on January 11, 2004 from http://www.teradata.com/t/pdf.aspx?a=83673&b=86857.

Carino, Kaufmann, & Kostamaa (2000). *Are you ready for yottabytes? Storehouse Federated and object/relational solution.* Retrieved on January 10, 2004 from http://siteseer.nj.nec.com/carino00are.html.

Inmon, W. H. (1996). *Building the data warehouse* (2nd ed.).Wiley Computer Publishing New York, NY.

Strange, K. (2003). *Data warehouse DBMS magic quadrant: Battle intensifies.* Gartner Ras Core Research Note M-19-2009. Retrieved on January 11, 2004 from http://mediaproducts.gartner.com/gc/webletter/ncr/article18/article18.html

Figure 1

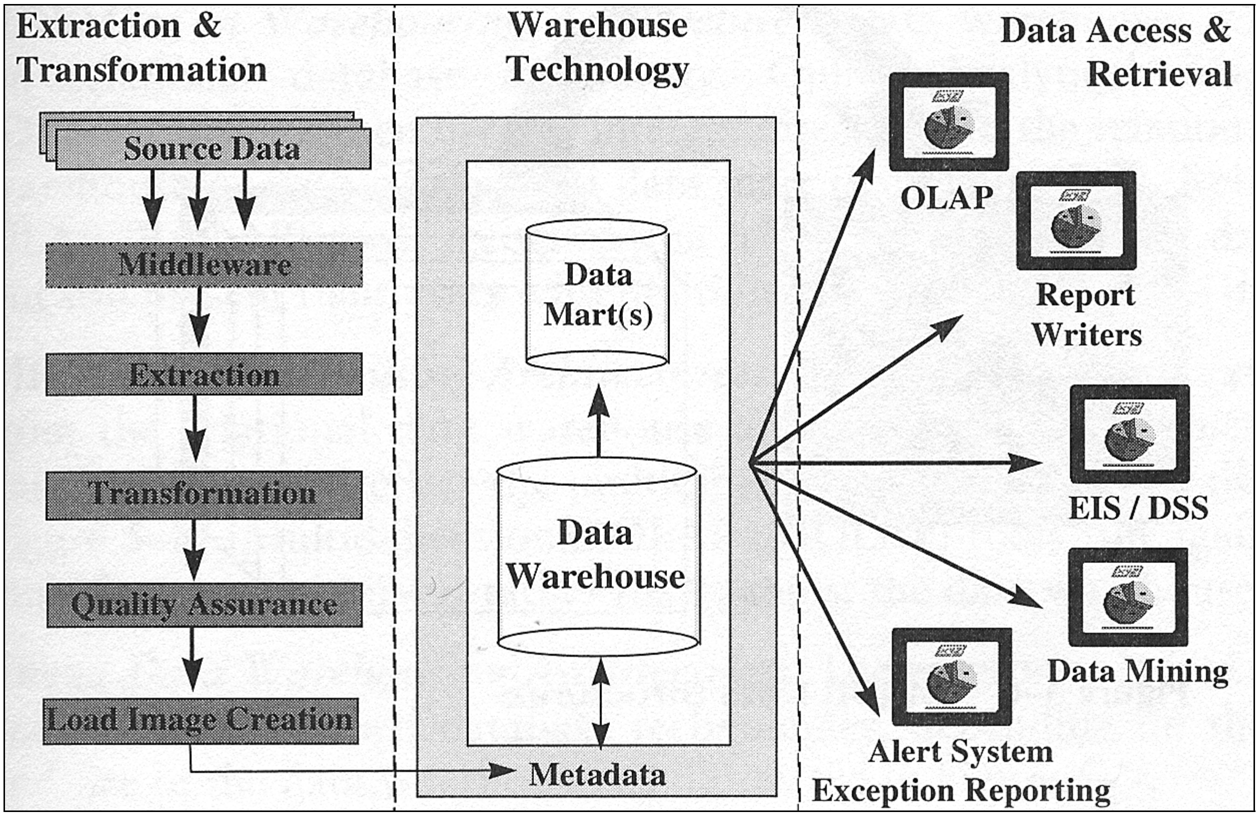

Figure 1. Warehouse Access and Retrieval Demands (Humphries, Hawkins, & Dy, 1999, p. 111)

Figure 2

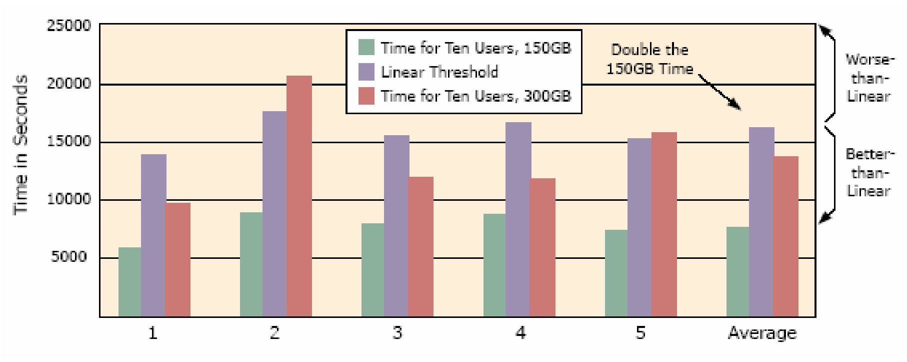

Figure 2. Linier scalability as data volumes grow (Ballinger, 2003, p. 8)

Figure 3

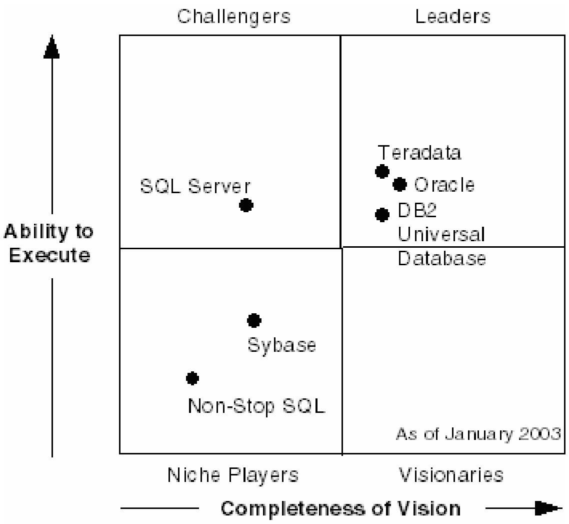

Figure 3. Data warehouse database engines (Strange, 2003, p. 1)