IN MEMORY ANALYTICS: BUSINESS INTELLIGENCE WITHOUT A DATAWAREHOUSE

Anthony Serapiglia, Robert Morris University, serapigliaa@rmu.edu

==

ABSTRACT

Thirty years ago System Memory (RAM) was expensive and processors were slow. Faced with these constraints at that time, developers devised an architectural approach for delivering results of multidimensional analysis which relied on pre-calculating fixed measures. To facilitate this, the answer was to physically separate the transactional databases from Data Warehouses. This action was driven by the need to avoid performance degradation of the transactional databases as complex queries were run against it. The difference in price and performance today is astounding; memory is much cheaper and processors are much faster. Systems are being developed that will allow for the real time querying of live transactional databases without performance loss. This allows for real time Business Intelligence, without a Data Warehouse. This paper is a case study of one small business environment and their transition to utilizing In memory Analytics through more powerful hardware.

Keywords: In Memory Analytics, Business Intelligence, Real Time Reporting, Data Warehousing, 64Bit, High Memory Systems

INTRODUCTION

Necessity may be the mother of invention, but cost and limitations of materials are ultimately the architects of any endeavor. So it has been with the history and development of the Data Warehouse. While many lay claim to being the father of Data Warehousing, such as Bill Inmon, Barry Devlin, and Martyn Jones (Hayes, 2002), by the time of the early 1980's one major architectural decision was already in place. The foundations of the Data Warehouse were created out of projects at MIT during the 1970's that worked on a wide array of technical architecture issues. It was a time of change, of looking differently at data processing and seeing instead information management. But it was a time driven by technology of that moment, and times have changed.

The MIT projects laid a ground work that has been the foundation of much of past 30 years. Of the more significant, for the first time, the researchers differentiated between operational systems and analytic applications (Haisten, 1999). The intent was to develop architectural guidelines for developing new solutions from the ground up. A core principle that emerged was to segregate operational and analytic processing into layers with independent data stores and radically different design principles. One of the main contributing factors in this decision was the extremely limited processing and storage capacities that existed at the time which strongly motivated the desire to off-load the new and wildly unpredictable analytic demand from the transactional systems platform.

Once separated physically, both online/transactional database and data warehousing systems both evolved along separate paths philosophically. Data structures and schemas were optimized on both ends for the tasks that each had to perform. Again, these evolutionary developments were driven by a core factor of performance and the capabilities of the hardware driving the systems. Only so much processing power existed, only so much system memory was available. To be able to perform complex queries and Business Intelligence (BI) processing on a live transaction system without bringing performance to its knees was unthinkable. For most business, this necessitated the building and maintaining of two separate hardware systems. One for transactional databases and another for BI needs specifically. This has been a very costly scenario, with current pricing of full DB systems including hardware and software ranging from $500,000 for Microsoft SQL server to over and estimated $3 million for an Oracle database system (Microsoft, 2010).

During the first decade of the twenty first century, great advances were made in the capabilities of computer hardware. Coupled with the rise in performance was also a distinct drop in overall cost. These two changes were most distinctly felt in areas of processors and system memory. In the world of computer processors, by the end of the 2000 decade 64 bit, multi-core processors were commonplace not just in server equipment, but also personal computing devices. Cost of a representative 64 bit processor in

2005 was approximately $1,161 (Kashi, 2005). A representative 64 bit processor of 2010 could be purchased for under $500. Single system motherboards are available in the consumer market that can hold upwards of 196GB of RAM. The price of RAM, once steady at a dollar a MB, has found a new stable range at less than three cents a MB (New Egg, 2010). These prices make "Super" systems very much within the reach of even the small to mid-sized business marketplace (SMB).

While these two areas of hardware have seen improvement on performance and speed, a third area related to database performance has seen a near revolution leading to almost obsolescence of the status quo. The traditional philosophy of storage has been radically shifted away from mechanical spindle driven hard drives to an all electronic solid state driven memory solution. Flash-based Solid-State Disks (SSDs), and DRAM-based SSDs have grown in size and popularity as their prices have fallen. Their use as primary storage has become almost mandatory for any at load DB system, as seen by recommendation by IBM, Sun, and EMC (Shu, 2009; Whitehorn, 2009)

High-transaction databases are typically comprised of small records (i.e., 4 or 8 Kbytes) that are often accessed randomly. Because the records are brief when compared to the time required to reach the data location, mechanical disk drives are paced by their ability to locate and retrieve information on the disks (i.e., disk access time). Disk access time becomes the dominant reason for slow database application performance, often causing the CPU to wait for disk I/O to complete. The problem is compounded by ever-increasing disk drive capacities because, as companies take advantage of larger drives, they are accessing fewer spindles for the same amount of data. With just a 50 GB drive, less than 0.004% of the drive is accessible each second when randomly accessed for database-sized records. Striping information over multiple drives can increase the amount of data that is accessed, but transaction integrity requirements of the database often conflict with broad, simultaneous access.

Elimination of performance bottle necks has been achieved on multiple levels, and in dramatic fashion since the MIT experiments of the 1970s. Processors continued to follow Moore's Law (Moore, 1965) in growth and the advent of consumer 64 bit technology has fostered that growth even further. System, or RAM, memory has also increased in speed and reliability with the cost become a small fraction of the previous norms. Through use of the latest solid state drive technologies that move past mechanical devices and rely solely on chip technology, IO storage factors have also been tamed. All of this gain in performance has to eventually lead back to the original decisions to separate BI analytic needs from the transaction data collection and serving needs. Does this separation still have to exist by necessity? Or can real Business Intelligence be conducted on a transactional DB without bringing performance to its knees? The purpose of this paper is to detail one company who has been able to achieve this pairing of the two sides in one system and really can have valuable, real time BI without the data warehouse.

COMPANY A

Company A, LLC operates as a scrap tire collector and recycler. It offers used tire collection, processing, and disposal services in the United States. It serves private retail tire dealers and government agencies. The company also sells shredded tire chips for tire derived fuel, civil engineering, and ground rubber applications. Company A, LLC was incorporated in 2002 and is based in Pittsburgh, Pennsylvania. Company A has a nationwide network of 20 production facilities. The company collects and recycles about a third of the nation's scrap tires.

As a result of shredding and grinding more than 1 million tons of scrap tires annually, Company A produces 1.5 billion pounds of recycled rubber in various sizes that are incorporated into advanced rubber-based products such as artificial sports fields and tracks, rubber mulch, road asphalt, railroad ties, auto parts and molded goods.

"Company A expanded prudently with an eye toward bringing an economy of scale to the tire recycling business - to build on the best practices of the companies we acquire by making capital available to them," CEO Jeffrey Kendall says. "Pursuing that vision, we have helped transform the industry. In the recent past, nearly 90 percent of the tires Americans consumed entered the waste stream and only 10 percent were recycled. Now, nearly 90 percent are recycled and only 10 percent enter the waste stream (Kendall, 2010)"

With the largest network of tire recycling facilities in the nation, Company A provides one-stop, coast-to-coast tire collection services at more than 60,000 locations. Tires of every shape and size are collected from a vast line-up of customers. The company maintains a nationwide network of door-to-door, comprehensive reclamation services and processing plants at strategic locations throughout the country.

Collection services performed by Company A include back-door pickup, whereby employees equipped with a fleet of box vans visit customer sites directly. The company also provides drop-and-hook pickup. Customers who choose drop-and-hook pickup are left with bulk trailers to fill at their convenience, and Company A takes the load away with a tractor when full.

Processing starts with the destruction of tires, which Company A accomplishes in one of two ways. Mechanical systems shred and grindscrap tires into chips or small particles using an ambient process. And cryogenic systems freeze tires at extremely low temperatures, easily shattering them to create a variety of chip sizes.

The transformation of an industry that Company A was able to pull off relates directly to their integration of information systems and data mining techniques previously unseen in the trade. Many of the acquisitions fell squarely into the small business category, most having less than 20 employees. These smaller business brought territory, existing contracts and routes with wholesalers, retailers, and even military bases. What the smaller acquisitions did not bring to the deal was IT infrastructure. For the smaller companies, margins were slim. An investment in computer systems, accounting software, and qualified personnel to manage these areas was an impossible scenario to bring to fruition.

With reasonable capitalization, and an existing Information Systems structure that was current and well managed, the smaller properties were able to be assimilated into the larger umbrella organization easily and the data accumulated from their operations quickly processed to find efficiencies in route management, disposal methods, and allocation of resources. The greater Business Intelligence process created a profitable organization that has grown from $35.5 million in revenue in 2005 to 110.6 million in 2009 (INC, 2010).

THE SYSTEM

At the beginning of the process of acquisition and growth, Company was truly a Small/Mid-sized Business (SMB) itself. With employment of 89 in 2004, only 24 of these workers were in the main business office. The Company oversaw only three production facilities at this time. MSSQL2000 was the main data repository for a system running a legacy management product that was 5 years old and no longer supported by the software vendor. This was soon replaced with a newer software package designed by the manufacture of the large truck scales that were in use at the three facilities, PC Scales. This package was then coupled with a SAGE MAS 500 ERP solution. For the next 4 years this system provided enough of a foundation to effectively run the company though its early growth period (Figure 1).

Quickly, as business grew, this system became sluggish and even prone to breakdown. Data pump mechanisms were used to extract figures into preset reports. Often these batch jobs would fail, requiring them to be cleared and re-run from the start. Several turnovers in personnel left many of the reports static, with no one person knowing exactly how to alter them and necessitating time consuming efforts to create new ones. As the server machinery housing the database came to end of life, a new and more flexible plan was put in place. Two choices were available; build two separate database systems with one dedicated towards data collection and the other a warehouse devoted to analytics. The second option was a more powerful solo system that would be able to run developing software that would enable both tasks to live simultaneously on the same system. After consulting with software vendors, infrastructure managers, and their own accountants, Company A decided to go down the path of a single system.

Taking advantage of the prevailing hardware environment, a new system was developed that included dual multi core 64 bit Intel processors, at 3Ghz. Running Microsoft Windows 2003 Server at 64bit allowed for 196GB of RAM to be installed. SSD drives were utilized to create a multidisc RAID 10 array. Tested IO speed reached sequential read 220mb/s and write 200mb/s. Microsoft SQL 2008 (also 64bit) was utilized as the Database Server software. This system cost Company A approximately 40% less than the projected cost of a multiple server / warehousing alternative system (Kendall, 2010).

Data is placed into the system through multiple real time entry points. PC Scale software takes readings from truck scales on location and forwards the data over a VPN connection for direct entry into the corporate DB in Pittsburgh. There are currently 20 such locations tied directly to the main office. Route Track software also feeds directly to the corporate office. This system collects GPS information from the fleet of trucks that run routes for pickup and delivery. At several key production facilities, newer shredder/grinder machinery also feeds detailed data directly into the system in real time. Currently, first quarter 2010, the main database size is 125GB. The entirety of which is able to be held completely in system memory, or RAM.

Each individual site has the ability to run reports ad hoc through the SAGE MAS software. These are constructed and produced while logged into terminal server sessions at the main office. IP printers mapped back to the individual location allow for the printing of hard copies at any location. Sage reporting ties directly into the live DB to pull information on demand. This is possible through the use of Sage SalesLogix Visual Analyzer (Figure 2).

Sage SalesLogix Visual Analyzer is built on a simplified architectural premise that all data should be held in memory, and all calculations should be performed when requested and not prior. The SalesLogix Visual Analyzer solution is and advantage over traditional OLAP and static reporting solutions because of this ability to pull data entirely from memory. Traditional OLAP solutions are pre selected and not flexible to easily add measures outside of the original cube.

Company A's managers have the ability to add measures on the fly to truly create a real time window into the activities of their entire operation. Every piece of data is housed in one location and available at any time. Dashboards can be created and altered to display graphically the performance of processing or collection centers. Their readouts can be automatically refreshed to instantaneously reflect the precise levels of the moment. "My day, and a lot of other's around here, were built around when the numbers came in and when we would be able to get our hands on reports and figures from the previous day or even week," relates CEO Kendall. "Now I can see at a glance, at any time, what is going on in any of our plants - if one has bad numbers, or if one is overachieving. Some might call it micro-managing. I call it keeping my finger on the pulse of the company. My meetings are not called only when I finally get the data, I can now proactively schedule anything knowing I'll have exactly the information I need when I need it."

CONCLUSION

With revenue now expected to top $200 million in 2010, and employees numbering over 650 in 20 locations, Company A has grown out of their SMB beginnings. Yet they are far from being considered a Fortune 100 company yet.

The business plan has been very sound. To take up smaller entities in the market place that could not invest in their own infrastructure, place solid Business Intelligence technologies in place, and assimilate into a greater whole to find efficiencies and leverages to build the company stronger. This strategy just would not be possible if it were not for the ability of the data systems supporting it. Through real time data analysis, facilitated by in memory analytics performed on the live transactional database, Company A has been able to efficiently make business decisions quickly and accurately.

In taking advantage of new, cutting edge hardware technologies, Company A has realized an initial setup savings of at least 40% over an alternative warehouse system. The new "In Memory" system has allowed for extreme growth and flexibility. The intuitive interface of the Sage SalesLogix BI tool has allowed more managers and accountants access than ever before giving a dynamic advantage to decision making previously impossible.

This study has illustrated one example of a company that has been able to take advantage of prevailing industry costs and availability of hardware and software. The focus of this paper has been to illustrate the structural shift in the processing of data into information, in the rearranging and eliminating of what used to be held as standard parts of a business intelligence system. Further research should be focused on sever other aspects of this system and others like it. The success of the life cycle of these systems needs to be studied as they grow and the businesses around them continue to add more and more data into their databases. Usability studies should also be conducted of the data analysists who are free to run customized reports at any time. It is an assumption to conclude that this is a good and efficient feature. It may be shown through qualitative inquiry that the free inquiry is not as productive as structured cubes for some people or businesses.

Business Intelligence without the data warehouse is not just a novelty. It is fast becoming a necessity. With the power and availability of the machinery necessary to support "In Memory" analytics now in the reach many SMB's, it only a matter of time before anyone who wants to keep up has to implement their own or be left behind.

REFERENCES

Computer Hardware: Server Memory. (n.d.). Retrieved April 2, 2010, from New Egg: http://www.newegg.com/Product/ProductList.aspx?Submit=ENE&N=2010170541%201052307858&name=1GB

Haisten, M. (1999, June 15). *The Next Stage in Data Warehouse Evolution, part 1*. Retrieved March 22, 2010, from Information Management Online: http://www.information-management.com/news/946-1.html

Hayes, F. (2002, April 15). *The Story So Far*. Retrieved March 22, 2010, from ComputerWorld: http://www.computerworld.com/s/article/70102/The_Story_So_Far?taxonomyId=009

INC. 500 Company Profile: Liberty Tire Recycling. (n.d.). Retrieved February 20, 2010, from Inc.com: http://www.inc.com/inc5000/2009/company-profile.html?id=200915080

Kashi, J. L. (2005, February). *Is 64-Bit Computing Worth It? A Performance and Cost Comparison* . Retrieved April 5, 2010, from Law Practice Today: http://www.abanet.org/lpm/lpt/articles/tch02051.html

Kendall, J. (2010, March 2). (A. Serapiglia, Interviewer) Pittsburgh, PA.

Moore, Gordon E. (April 19, 1965). “Cramming more components onto integrated circuits”. *Electronics Magazine*, Volume 38, Number 8, Retrieved October 11, 2008 ftp://download.intel.com/museum/Moores_Law/Articles-press_Releases/Gordon_Moore_1965_Article.pdf

Shu, C. (2009, March 11). *A Performance Study of Using SSDs in IBM DB2 Applications*. Retrieved March 20, 2010, from Sun Systems: http://blogs.sun.com/cshu/entry/a_performance_study_of_using

SQL Server Cost Savings Calculator. (2009, October 15). Retrieved April 5, 2010, from Microsoft SQL Server 2008 Home: http://www.microsoft.com/sqlserver/2008/en/us/costsavingscalc.aspx

Whitehorn, M. (2009, August 24). *In the spin of SSDs on database servers, The future is static*. Retrieved 20 2010, March, from The register: http://www.theregister.co.uk/2009/08/24/whitehorn_ssds_servers

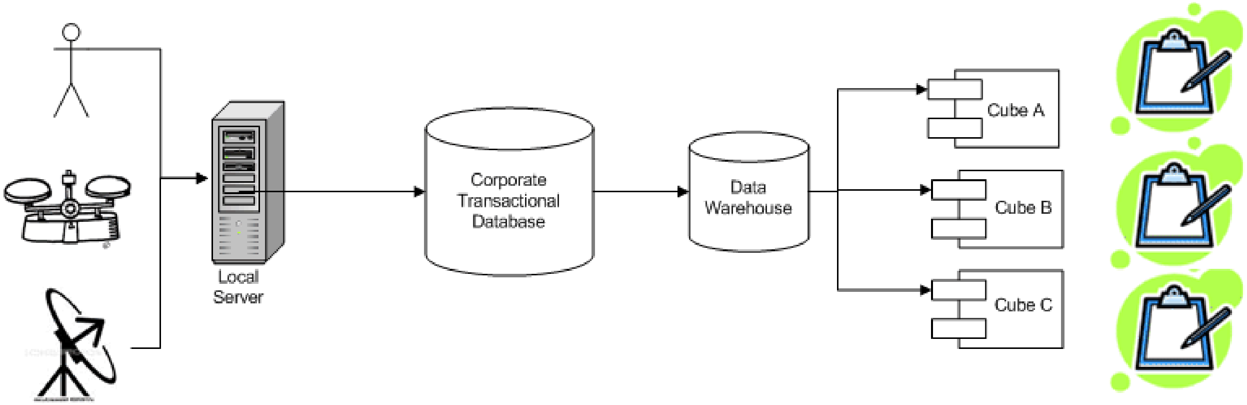

Figure 1: A diagram of the original, more traditional system. Multiple forms of input include GPS entry, truck scales, and manual entry. Data is initially collected locally, at each site, then uploaded into a central collection point. Data is then extracted to a data warehouse. Pre-designed cubes are constructed based on reporting needs.

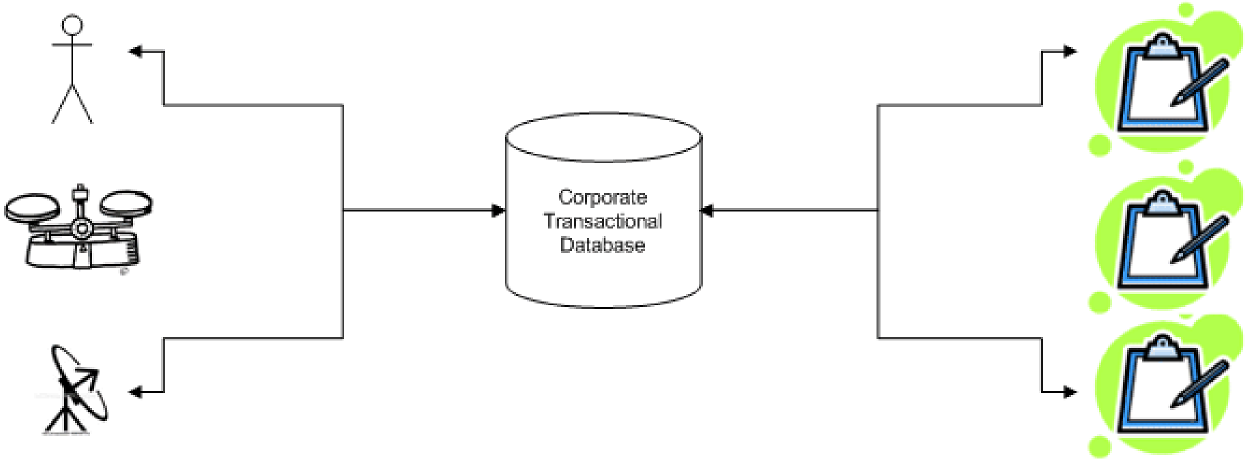

Figure 2: A diagram of the new, evolved system. Inputs from various sources are sent directly to the main collection point without a local stop. This central server is then queried directly as reports are generated ad hoc.