

The Enterprise Data Cloud: Leveraging the New Economics of Data

Merv Adrian, Principal, IT Market Strategy
www.itmarketstrategy.com

Executive Summary

Rising IT costs, exploding data volumes, and ever-evolving competitive challenges have catalyzed new ways of thinking about effective systems architecture. In recent years, leading organizations have rethought the way they assess, procure and operate the hardware assets they use to manage data. At the same time, new programming techniques have transformed the skills and methodologies innovative firms apply to analytics for business optimization. Both sets of changes have had driven radical changes in database technology, and collectively, these developments have led to a new approach for effective data exploitation that is being called the Enterprise Data Cloud™. In this paper, several leading edge practitioners describe their operating environments, and we describe the framework and its benefits.

Introduction – Today’s Landscape Appears Difficult to Navigate

The need for agile organizational response to market conditions has never been greater, as globalization brings new competitors on what seems an almost daily basis, from anywhere in the world. Worse, these upstarts have access to data streams and techniques for analyzing them that used to be the exclusive province of the largest companies, with the biggest computers and armies of technology specialists. The barriers to entry are lowering, and many established organizations feel as much bounded by their legacy systems and data as empowered by them – the costs of managing their legacy systems are getting in the way of staying ahead of growth.

Exploiting information has become an imperative for all businesses, and its importance – and the opportunities - grow daily as rates of data growth accelerate. But the costs of proprietary servers and storage devices, space and the energy to manage them are off the charts and highly visible to every CFO, CTO and procurement professional. “More of the same” will not work. Proliferating copies of data into multiple one-off analytical systems – seemingly one for every question to be asked - only adds to the costs, and even the new “data appliances” can cost in the tens of millions to scale up as needs grow, compounding the problem further.

Moreover, hidden costs – in people, processes, and performance – are at least as significant. People with the specialized skills needed are hard to find and retain. Keeping multiple data copies synchronized and compliant with privacy and regulatory requirements is difficult and time-consuming. Meeting performance needs that are driven by demanding users seeking innovation is a constant battle that absorbs money and time in equal measure. For all its promise, the data explosion exacerbates the difficulties of achieving business exploitation of

information in a cost-effective way. Applying the lessons waiting to be learned inside their massive data stores seems out of reach to CIOs who don't see how they can live within their budgets.

Hardware and Software Economics Challenge Budgets

Data warehouses grow fast. If you haven't looked at the state of the art for a few years, you might be surprised to hear how frequently terabytes – even petabytes - are mentioned, even for firms of modest size. Consider Zions Bancorporation, a mid-sized financial institution operating in 10 Western U.S. states and growing, with frequent acquisitions. Their data warehouse, operating for about 5 years, has reached the 6 Tb mark, and recently they found that its existing Sun server and Oracle database – a widely used combination – were in need of an upgrade. Soon.

There was no time to waste. Zions Bancorporation was struggling with their ability to load data, and the performance of night batches was starting to impact their ability to meet service level agreements with their constituency. And once the data was loaded, query response speeds were slow. “We boiled the solution down to a few possibilities,” says Clint Johnson, Vice President of Business Intelligence, “and all involved serious money.”

A big piece of the challenge was hardware. The warehouse was running on Sun Fire 4800s, with data loaded into a storage area network, and throughput was becoming a problem as Zions Bancorporation hit IO challenges. The costs of conventional data warehouses do not scale linearly to volume – they're “lumpy,” as new server and storage capacity are acquired in big chunks. Vendor pricing models vary, adding complexity: “Even when supplier bids got close, the marginal cost of new capacity was measured very differently,” says Johnson.

The hardware to run Oracle, for example, needed to leap to a next-level server and dramatically more storage just to keep up with the requirements Zions Bancorporation was already able to foresee. And when the next acquisition caused a substantial jump in volumes and numbers of users, that hard-to-predict spike could make the increment even greater. The newly acquired servers, storage, interconnects and software might not be compatible. Migration costs would add to the total, and skills might be needed that Zions Bank did not have. In addition, the “data tax¹” made it likely that software licenses tied to hardware and/or user volumes would leap.

Appliances Attempt to Redefine Hardware Economics

One hardware alternative Zions considered: proprietary MPP appliances. Vendors like Teradata and Netezza both offer proprietary models that have clearly defined – though large – scaling steps, and at least servers and storage could be considered together.

In 2002, the first “data appliance” – computing plus storage - was introduced. Netezza Performance Server configurations were offered in three sizes: 4.5TB, 9 TB and 18 TB, starting at \$622,000, a breakthrough at the time. The premise was straightforward - hardware, software and storage in one purchase, preconfigured and certified to work with the BI products customers were using.

¹ Many software products are licensed in part by processor size, or by the number of cores used. Since more data typically demands more processing power, data growth often leads to higher software costs.

In a market where large scale specialty machines for analytic use were only being marketed by Teradata, this pricing was a shock. As the idea caught on, other vendors began to adjust and create their own versions of the form factor. IBM leveraged its vertical integration with the InfoSphere Warehouse running on its own servers and storage. The new model seemed to attract buyers, and in the next few years appliances grew their share of market steadily.

After the possibilities were considered, however, these closed models ultimately created challenges Zions did not want to take on. “We were comfortable with Sun, and we like the ability to add hardware easily,” says Johnson. “Netezza was more specialized and scaled in very specific size increments, and Teradata required a high level of support and services that would have required us to add skills to manage ourselves.”

Software issues also loomed. Teradata licensing costs scaled with total combined capacity, so even if over-provisioning hardware was acceptable, it brought added software costs too. Similarly, Oracle RAC, another architectural possibility, brought “licensing that was dramatically higher than the conventional databases licenses we had from them.” Zions Bancorporation concluded that they needed to find a different model, one that used commodity hardware that would mix and match, without special interconnects and administration needs, and with software licensing that scaled cleanly with the hardware scaling.

Such opportunities do now exist. In recent years, Oracle has created Optimized Warehouses with pre-installed database software to be marketed with its hardware partners. DATAlegro (now part of Microsoft), Dataupia, Greenplum, Kognitio, Paracel, Vertica, and others have also emerged to offer lower-priced software offerings running on commodity hardware. As Mark Dunlap, a consultant with Evergreen Technologies and a veteran of massive data warehouse projects at Amazon and Fox Interactive, puts it, “If you're using proprietary hardware, you're in a losing battle. Sooner or later, whatever company's developing that technology will not be able to keep pace. We've seen it over and over again - they won't keep pace with what commodity systems are doing.” The “commodity hardware” model quickly challenged the leaders for mindshare in smaller, analytic application opportunities.

However, the proliferation of small appliances has begun to emerge as a new problem in its own right. It has created chaos as small, admittedly nimble and useful installations were purchased by line of business executives who did not always coordinate with IT, content to solve their immediate problems in isolation. Moreover, existing appliances missed an additional key opportunity in commoditized hardware: capacity optimization. They were typically underutilized, one application at a time, not flexibly provisioned across hardware instances.

The problem here again ought to be simple economics – unused capacity is a needless expense. Organizations need to make the most use of their investment in hardware and storage, and their software needs to keep up. But measurements of capacity utilization in most shops show that as much as three-quarters of system power is not in use at a given point in time. So, even with enterprise scale resources, organizations retreat from tackling enterprise-scale problems. Disconnected appliances bound the solution set, and draw barriers that constrain the art of the possible. Harnessing compute power effectively by treating it as an available resource pool that analysts can draw on at will, and providing access to data wherever it resides, will give organizations the power to tackle problems they never thought possible.

BI Requires Provisioning On Demand

As IT executives charged with reducing costs, space and energy usage redouble their efforts to manage server resources better, storage managers and data stewards face similar issues: in a 2008 poll, 62% of 136 IT leaders from the U.S. and Canada called resource management their biggest challenge.² What drives this problem is the inexorable acceleration of data growth, and our need to exploit it. The “one server per analytic application” phenomenon discussed above has created server sprawl, as “spreadmarts” have proliferated across the landscape. There is a good reason for that: there are always more questions, important ones, to ask. Empowering analysts to ask and answer them is a critical imperative.

Telecom analysts know their users better than marketers in most industries. At T-Mobile, “people use their handsets more and more, for voice and data, and we track that usage,” says Ryan Hawk, Director of Information Management. Trending is key, but trends change constantly. Hawk’s team builds models – propensity to churn, revenue generation, and more. Complex techniques, statistical analyses, and new analytical models emerge and disappear. All need compute capacity and data – but at these volumes, “data is a business case – we have to decide what we can afford to store on our MPP systems,” Hawk says. “The hardest thing is having to purge data every 60 days – you can’t do much trending. With more granularity, you get into what happened, when and why.” T-Mobile needs more data, for more flexibility. And more capacity on demand, to analyze and rethink. They need “sandboxes” – build a data set, analyze. Discover new questions. Grab data to explore those questions. Rebuild. Repeat.

The appliance model only goes halfway to the solution: while processors and storage are coupled for effective scaling, they can’t be decoupled easily to scale a different problem than they were first designed to solve. Easily provisioned storage and servers are the critical enablers, facilitating iterative buildup and teardown of analytic projects. BI applications of virtualization using inexpensive, multi-core hardware will liberate pent-up analytical teams: they can exploit new, flexible provisioning models built on commodity hardware that can be scaled at lower cost. Hardware performance is accelerating; parallel software that exploits it has lagged. No more.

This more effective compute resource paradigm decouples individual problems from individual servers (or cores). As projects emerge, a virtualized system “flexes” – grabbing unused resources and lashing them together until the computation is complete, the process has been executed, the prediction composed. And all this can be done without anticipatory procurement designed for the highest possible momentary peak – today’s all-too-frequent, costly approach.

There is a strong correlation between this flexible provisioning and sharing of resources and the recombinant, exploratory nature of true BI. Analysts restricted to carefully designed data structures, in rigidly cubed, indexed and aggregated forms designed for known questions, resemble the man looking for the keys he dropped near his car under a streetlight half a block away “because the light is better here.” They limit the questions to be asked. Insights can’t be predicted, and known analyses of known data give predictable results that are confined to problems we already understand. New programming approaches unconfined by old designs – and even existing programming languages – are evolving to tackle the new data flood.

² Storage Magazine, April 2008 survey conducted by the Excillio Group

The Enterprise Data Cloud – Redefining Analytic Architecture

The label “Enterprise Data Cloud” is being popularized by Greenplum, a parallel DBMS vendor whose architects and engineers have participated in the academic and commercial development of many of the recent architectural innovations described above. Their customers, some of whom were interviewed for this paper, have pushed Greenplum to drive their Postgres-based engine to support and exploit the enabling technologies we have described. Two core innovations will be required to achieve all the promise of this new system blueprint: to exploit commodity hardware provisioning, we must add parallel database capability that supports not just multicore exploitation, but also easy creation and distribution of instances, easy or even automated resizing, and new techniques for incorporating external data sources.

The bottom layer is hardware infrastructure. It presumes virtualization – for the user of the system, system resources need to be available on-demand, flexing as needed to accommodate more server capacity, for example. Brian Dolan, Director of Research Analytics at the Fox Audience Network, says, “I get to share 40 nodes with the production system. I use them when I need them, and then I give them back.” The data cloud will be primarily targeted at creating an internal structure for building “sandboxes” as needed – mapping servers (or cores) and data stores into the form needed to address the task at hand. The exploitation of public clouds like Amazon’s can extend the model to incorporate data not owned by the enterprise but useful for rich model creation.

Changes to conventional database practices are critical enablers. An architecture that truly supports parallelism in shared-nothing fashion needs many operational tools to be in place. Well-managed identity management and authorization are vital, to permit analysts to extract data from multiple sources, as well create and tear down new data stores as on-the-fly. The management of data lineage is more complex in such an environment, as is replication for disaster recovery and availability. These operational aspects must be seamless and not intrude upon the users of data, but controls are needed for administration purposes – managing users, provisioning instances and moving data as needed. It takes good controls to liberate usage.

Where the action is, the system must manage parallelism (execution and data distribution) across nodes, control the interconnect’s operations, optimize parallel queries and create materialized datasets on demand. In-memory, columnar and more conventional relational persistence models may be appropriate at different stages of complex operations, depending on update frequency, the nature of the operations being performed, and the volumes of data. Similarly, with multiple storage options available, solid-state memory, high speed disk and less expensive storage media will all play a role and their most cost-effective use will need to be determined. The ideal system will support them as well. One last component remains that the system must support: sophisticated in-database functions for extensibility into new programming approaches.

Analytics Programming Paradigms Evolve

For decades, analytical applications have obtained data primarily by invoking SQL database engines to retrieve and manipulate it. Today, virtually every BI product uses interfaces such as ODBC and JDBC to retrieve and pre-process data they need; then more work is done on the extracts - somewhere.

But this approach falters in the face of the massive volumes some users require. Brian Dolan at Fox says, “Groups I’ve worked in exported data into local machines to run sophisticated analysis. But some problems could not be processed there. For example, I asked SAS whether they could perform certain analyses over an array of the size we were dealing with, and they said they could not.” Dolan’s team might work with 2 weeks’ worth of data: 100 billion lines, 10s of terabytes. Exporting, transforming, moving and distributing in chunks (extract, transform and load: ETL style), constrained by bandwidth and system load factors can take 3-4 days, rebuilding all the joins and the structure within the data, another one or two. So 5 or 6 days later you can actually begin using it. Not good enough, suggesting the question: *what if you could have the provisioning you needed for on-demand creation, and a database platform that could use those resources to scale? What else would you need?*

In recent years, extensions to SQL, and the addition of callable functions within database engines, have enhanced some DBMSs, allowing some problems to be tackled without large exports and new file creation. For Dolan, using clever SQL joins and table structures offered a solution to his problems, and it got him part of the way there. “We simply don’t have the capacity to analyze the data except within the database,” he says. “Inside the database, setting up that two weeks’ worth of data takes us about 20 minutes. Generally if we’re doing it after 11 in the morning we’re not in anyone’s way – and we can look at the system to ensure there isn’t a lot of stuff going on,” Dolan says.

For other needed tasks, additional methods have emerged. The MapReduce programming model pioneered by Google has captured the attention of many developers. It’s very valuable in dealing with external files, rapidly extracting needed data from massive volumes of records – and without traditional database programming skills. Supporting MapReduce scripting in the DBMS can extend these new methods to enable external data use to supplement existing structured data. This allows programmers to write transformation scripts in the dataflow-style programming used by many ETL tools, while running at scale using the DBMS’ facilities for parallelism. A recent academic paper describes the alternatives: “the choice between ETL and ELT is sometimes a trade-off between filtering early (favoring ETL-style transformation close to the data source), and exploiting the power of a DBMS for complex, massively parallel data transformation (favoring ELT-style transformation in the target DBMS). Scatter/Gather into the database is critical in either case.”³

Three key enablers, then, come together to raise the bar. With lower cost commodity hardware offering power at costs an order of magnitude lower, on-demand provisioning allowing flexible, recombinant exploitation of system resources when needed, and DBMSs with the ability to exploit parallelism with sophisticated internal functions and support for new programming paradigms, the stage is set for a new design center: the Enterprise Data Cloud.

The Time To Innovate Is Now – The Price Is Right

Many IT innovations, especially at the high end of scale and value, are first seen in the largest firms with the IT resources and budget to experiment, often with large initial costs expected. The newest technology thus often changes fundamental elements in ways that provide a barrier to entry for those who can’t afford to play, and as a result, smaller businesses content themselves

³ MAD Skills: New Analysis Practices for Big Data, Hellerstein et. al. March 2009

with the idea that they will have to wait, along with everyone else, while the big guys get another opportunity to distance themselves.

Not this time. This opportunity is different – not only is the hardware a pricing leveler and thus accessible to a larger group of prospects, the software is not based on proprietary technology that carries a higher price tag than the alternatives. In fact, the entire package is cheaper, more extensible, and more easily pointed at different problems multiple times without new costs.

Organizations not yet pursuing an enterprise data cloud agenda should consider:

- Commodity hardware and parallelized software drive easier, more granular scalability, and compute and I/O advantages that improve system performance.
- Elastic self-service provisioning enables iterative analysis – answering the questions raised by the previous questions – to drive agility and flexibility both in IT, and if used well, in the business itself.
- Unified access to data for multiple programming models and easy incorporation of newer data streams and sources redefine the ability to analyze and improve business performance.

IT Market Strategy expects to see rapid adoption of this model. Your competitors may already be preparing – or piloting – an enterprise data cloud. Consider your plans for the next 18 months, and think carefully about whether this model might be cheaper, faster, and more agile than what you have in mind for one of your projects. It would be a shame to be left behind.