CLABBY ANALYTICS

Advisory

Business Analytics Buying Criteria: The Three Maxims

Introduction

At *Clabby Analytics*, we believe in three maxims (fundamental principles) when it comes to the acquisition and deployment of information systems that run business analytics workloads:

> ***First and foremost, no single microprocessor/systems environment can process all workloads optimally. Some designs excel at serial processing, others at data-intensive processing, still others at parallel processing. Choosing the right system architecture for your workload can save BIG MONEY.***
>
> ***Second, it is more cost effective to purchase an integrated infrastructure solution from a single vendor that offers an integrated infrastructure stack (systems software such as operating environments, virtualization hypervisors, and middleware) than it is to purchase piece parts from a bunch of different vendors and then attempt to integrate those parts under your own auspices. Some vendors have deep tuning expertise that can be used to integrate hardware, systems software, infrastructure, applications, and databases (these vendors have the ability to tune their components to the extreme). Most information technology (IT) organizations do not have internal access to this type of expertise.***
>
> ***And, Third, when deploying business analytics applications, failure to construct an infrastructure plan at the outset will likely result in scaling and performance complications that will be expensive to fix later.***

What leads us to these conclusions? Consider maxim #1. The reason it is important to choose the optimal processor/systems design, is related to cost. Running the right workloads on the right systems leads to greater performance and superior economics (fewer servers need to be purchased, licensing costs go down, management is simplified, networking is simplified, power and cooling costs decline, and so on). In a previous *Advisory* entitled "*Why Your Organization Should Use Workload Optimized Servers*", *Clabby Analytics* showed how the cost of computing differs on three processor/system architectures (mainframes, Power Systems, and x86 environments) running the exact same workload. This report shows how choosing the right server can result in *big savings*.

Regarding maxim #2, there are two reasons that it is important to purchase an integrated infrastructure from a single source vendor. First, infrastructure vendors have the ability to deeply integrate and tune their infrastructure components — saving IT buyers from having to spend time, effort and money melding together various software components. Second, purchasing an integrated infrastructure enables IT buyers to improve time-to-value (the time it takes to start reaping the value of an IT investment is greatly reduced).

As for maxim #3, it is important at the outset of a business analytics deployment, to select an integrated infrastructure that can support fast access by multiple organizations and individuals within an enterprise to data in various forms (structured and unstructured), as well as to allow access to potentially very large databases (Big Data databases). This integrated infrastructure needs to be tuned for performance to meet the users' service level

requirements. And it needs to allow for future scaling as the amount of data to be analyzed continues to increase.

> ***Failure to address performance and scalability at the outset of a business analytics deployment can result in systems and storage that underperform, network performance bottlenecks, and scalability issues during deployment (usually because the infrastructure was not designed to handle the volume of data to be analyzed). Another common outcome is having silos of departmental systems deployed, resulting in a fragmented approach to enterprise analytics, that delivers inconsistent, incomplete information to the business users. Redesigning an infrastructure at a later phase in a business analytics deployment can be time consuming and expensive.***

In this *Advisory*, *Clabby Analytics* looks more closely at the importance of implementing an integrated infrastructure — and its relationship to business analytics:

- We describe why deploying applications on a strong, integrated infrastructure is strategically vital (for cost and time-to-value reasons).
- We take a closer look at business analytics workloads (we describe six distinct workloads) and explain how these workloads put different demands on underlying systems and infrastructure.
- We provide examples of systems designs where performance has been accelerated through infrastructure integration and database tuning. And,
- We conclude with a recommendation that enterprises that are serious about building a comprehensive business analytics environment — and enterprises that want high performance and a scalable growth path — look closely at how IBM integrates its systems/storage/network/database/application infrastructure to achieve exponential improvements in performance.

Getting Started — Why Infrastructure Matters

Information systems infrastructure is the highway between an application and the services it needs to execute its tasks. This infrastructure includes systems software such as middleware for program-to-program communications, virtualization software to increase resource availability, and operating environments to control computer resource use.

Information systems infrastructure also applies to supporting subsystems such as networks and storage. Storage infrastructure needs to be tiered properly in order to reduce storage costs. These costs need to be balanced with the ability to access data and the ability to meet service level agreement (SLA) and Quality-of-Service (QoS) performance, reliability, availability, and security requirements. Network designs need to focus on reducing latency.

> ***The goal in infrastructure design is to allow applications and data to flow smoothly and efficiently across a given computing environment. To do this most optimally, IT managers need a solid understanding of all systems and infrastructure elements — including systems, storage, networks, databases, security programs, tuning parameters, system tuning parameters, database optimization practices — and so on.***

Why IT Buyers Should Purchase Integrated Infrastructure Solutions

At *Clabby Analytics*, we contend that most IT organizations are not capable of performing deep integration. Our reasons for this perspective are that:

1. Most enterprises do not have the deep expertise in-house needed to highly optimize systems/storage/network/database/application performance; and,
2. Many enterprises have cross-domain organizational uncertainty as to where the problem lies (systems, software, storage, network) when bottlenecks or other problems occur.

Vendors that make complete infrastructure stacks, on the other hand, have deep expertise as the people who built the infrastructure products work for them (you can't get deeper expertise than that); and vendors have a vested interest in working across functional groups to ensure optimum performance for a solution (because optimized performance leads to increased sales).

> ***IT buyers have a choice: 1) build their own infrastructure themselves and attempt to streamline performance across that infrastructure; or, 2) rely on vendors with deep expertise that can build an integrated infrastructure designed and optimized for performance. We suggest that the latter will yield orders of magnitude better performance than the former.***

A Closer Look at Analytics Workloads and Characteristics

Our first maxim stated that no single processor/system design can execute all workloads the most optimally. Therefore, IT executives need to analyze their applications to determine where those applications will run best.

The way we see it, there are six classes of business analytics applications — and each class puts different processing/data handling demands on underlying system and infrastructure. These classes are:

1. Deep analytics;
2. Operational analytics;
3. Pre-defined reporting;
4. Ad-hoc queries;
5. On-line analytical processing (OLAP); and,
6. Advanced analytics.

How These Analytics Applications Place Different Demands on Underlying Information Systems

A closer examination of the processing requirements of these business analytics applications shows that they serve different objectives, and place different demands on underlying systems, subsystems, and infrastructure:

- Some analytics applications involve analysis of very large databases while others involve analysis of small or modest databases.
- Some analytics workloads involve analysis of very large data stores in real-time. This data can be streamed data, structured and/or unstructured data (the industry calls these large collections of structured and/or unstructured data "Big Data"). Big Data users often need to analyze massive volumes of structured relational data and or unstructured data (such as audio or video files) in order to formulate insights using advanced analytics. The timeframe to execute is also considered

important for this type of workload — SLA requirements may require results in minutes not hours.

- Some business analytics workloads are designed around on-line analytic processing (OLAP) tasks, while others focus on obtaining results using on-line transaction processing (OLTP) data.
- Some business analytics tasks require rapid response times and currency of data that is necessary when information is integrated into the operational processes of a business — delivering real-time information to the front lines (such as in call center applications).

Based upon these differing requirements, the starting point for enterprises looking to deploy business analytics is to gain an understanding of the processing requirements of the given analytics application. By understanding your workloads and expected outcomes, your enterprise will be able to select the systems, infrastructure, and software that best suits its analytics processing needs.

Comparing Two Different Business Analytics Applications

To illustrate the differences in how analytics queries differ — and how these workloads put different demands on systems and infrastructure — a good place to start is by comparing two analytics applications that use different query approaches: a deep analytics processing requirements compared with an operational analytics environment.

- A new class of analytics (deep analytics) has started to evolve, thanks in part to IBM's ground-breaking Watson system. IBM's Watson is a workload optimized system design that uses a question & answer (known as "DeepQA") to conduct complex analytics. *Deep analytics queries* are usually complex queries that often involve sifting through large structured and unstructured data sets to get answers. In Watson's case, the data analyzed came from encyclopedias, dictionaries, thesauri, newswire articles, and literary works (in other words: semi-structured sources) — but someday Watson may also use unstructured data as a basis for analysis.

  Another defining characteristic of deep analytics is that it tends to involve low concurrency of queries (the processing very few queries as opposed to many queries simultaneously). This is because the types of queries being executed are so resource-demanding and have tended to take so long to execute, that it is not practical to process many of them at the same time. In Watson's case, the whole system was focused on seeking the likely answer to one query. The focus of deep analytics is to get an analytics result as quickly as possible — so there is a big focus on achieving rapid response time from the system running the query.

Deep analytics queries run best on systems that reduce the run time it takes to execute a query. Deep analytics IT buyers, therefore, should look closely at a given processor's ability to handle parallel processing. If the query processing can be parallelized, (separated into threads that execute individually and are then gathered together to produce a result), execution time for that query may be dramatically lower than executing it as a whole. This is commonly called a massively parallel processing (MPP) architecture — an approach that is common to systems that perform this type of work. The amount of data that can be placed in memory is also a key consideration as data in memory can be analyzed faster than data on disk.

- Operational analytics (this includes predictive analytics) involves processing *many concurrent queries from multiple users* — some that may be simple, while others may be complex — using a data warehouse back-end. In this case, lots of threads of execution are taking place simultaneously. The focus when processing operational analytics is to get a lot of queries done quickly — so *throughput* (how many queries per second can be executed) is a major focal point when processing operational analytics. *The number of queries being processed, the variety of the queries, and the number of queries that must be completed per second make operational analytics distinctly different from deep analytics queries*.

In deep analytics, it is all about finding and focusing resources on executing a single task against a large database. In operational analytics, it is all about getting many tasks done rapidly. In each case, the workload places different demands on a given system, subsystems and accompanying infrastructure. For instance, in the operational analytics environment, systems need to be able to dynamically reclaim and reallocate resources once a query has completed. Some information systems do this better than others. Mainframes, for instance, are a shared everything architecture designed to constantly return unused resources to a common resource pool. Systems design, therefore, plays an important role in business analytics application performance.

The Other Classes of Analytics

The other classes of analytics described earlier also make use of systems resources in different ways. For instance:

- Pre-defined reporting is often run in off peak hours during ETL (extract, transform, load) cycles — so the focus for this type of activity is to have reports executed within a given period of time — within guaranteed SLA windows (hence infrastructure QoS response is important to this type of query).;
- Ad-hoc queries are generally queries designed to validate various theories. The infrastructure challenge with these types of queries is that the syntax of the query statement does not exist ahead of time — making it difficult for administrators to tune databases to efficiently predict and process such queries. As a result, the time to execute a task may be lengthened.
- OLAP queries are organized into cubes with pre-aggregated information available to speed-up query processing. This type of query benefits from fast, multi-threaded processors, fast storage subsystems, and an infrastructure that can meet business requirements for resiliency and security. And finally,
- Advanced analytics encompasses multiple types of analytics workloads — including data mining and predictive analytics. This type of workload might run best on a hybrid architecture where different systems, such as mainframes and Power Systems and System x servers, all perform the work they are best suited to execute — and the result is achieved when the computing functions that have been completed by each system, are rolled-into a final answer set.

Evaluating Business Analytics Systems — What to Look For

When evaluating business analytics systems, IT buyers need to pay attention to:

- ***Processor characteristics*** — for business analytics environments, processors need to excel at processing compute-intensive and/or parallel tasks;
- ***Systems designs*** — some vendors offer pre-integrated systems designs that aim at executing specific types of queries. IBM also offers hybrid designs (where several different types of processors are used to execute various phases of a workload);
- ***Memory*** — some vendors design memory that is specifically optimized to serve their processors;
- ***Storage subsystems*** — are fast input/output per second (IOPS) required (if so, solid state disk drives better serve an IOPS-demanding environment than mechanical disks. Also, how data is stored (tiering) and data management are important;
- ***Data proximity*** — the closer the data can be placed to the processor, the faster that data can be analyzed. This means look at available memory, solid state disks, and on processor memory very closely; and,
- ***Integrated Infrastructure*** — the middleware, virtualization, and systems software components that expedite the delivery of data to the processor.

In short, your analytics design goal is to deploy your data in such a manner that it can be quickly moved into memory where it can be fed to the processor for execution. Processor capabilities, system design, memory and memory management, storage subsystems, data proximity and integrated infrastructure all play very important roles in business analytics systems designs.

A Business Analytics Systems Design Example: IBM's System z with a Netezza Appliance

Last October, *Clabby Analytics* partnered with *Enterprise Computing Advisors* to produce a report entitled "Chosing IBM zEnterprise for Next Gen Business Analytics Applications" which describes how IBM's System z (mainframe) processes analytics workloads. This report described how IT buyers can achieve superior economics using a mainframe configuration know as IBM's DB2 Analytics Accelerator.

For decades mainframes have been known for their ability to scale — and their ability to process large databases. The traditional approach has been to locate the data near the mainframe, install a lot of memory, and then execute an analytics workload as efficiently as possible. Last year, however, IBM improved on this design by adding a database appliance server based on Netezza, and modifying DB2 to recognize and route long running queries to the appliance. Netezza is an appliance that uses x86 and field programmable gate array (FPGA) processors to pre-process data. In this case, it can be used to filter very large databases and preprocess portions of the SQL. This, appliance can therefore dramatically reduce the processing time for queries. This effectively integrates an MPP architecture into the DB2 for a z/OS environment, allowing DB2 to selectively route queries to the appropriate architecture for the specific task.

First, this is an excellent example of a hybrid systems architecture. Second, the results of assigning the right workloads to the right processors have been spectacular — with some queries running up to 1908 times faster than the same queries running on a simple DB2 database server. This example shows why tuning systems and infrastructure is vitally important to achieve high levels of database performance.

Another Business Analytics Systems Design Example: IBM's System x with eXFlash

In January of this year we wrote a report entitled "IBM's System x with eXFlash: Workload Optimized Pre-Integrated Server/Storage" that described how an IBM System x server can exploit local solid state storage (contained in an 8 bay eXFlash storage cage) to yield exponential increases in performance. This report also described IBM's memory management subsystem that features a distinct memory controller (the current generation is known as eX5) that uses a dedicated application specific integrated circuit (ASIC) that off-loads the x86 CPU from having to handle a variety of communications tasks (for instance, the management of virtual server connections). Notice how this, also, is a hybrid design (it uses different processors to execute various tasks, just as the System z/Netezza example in the previous subsection did). Further, notice that it locates the storage very close to the processor (the storage is so close and so fast, in fact, that it acts like extended memory).

> ***This design offers up to a 30x improvement in local database performance — and is ideal for a variety of business analytics workloads including applications where real-time data can be stored on flash drives, where it can be quickly referenced — enabling purchase decisions to be made rapidly; surveillance and security applications where governments can run real-time security checks against reference materials; and data mining and database application environments that can benefit from solid state fast IOPS performance.***

The Newest Example of an Optimized System Design: IBM's PureSystems

The industry's best example of how a vendor integrates its infrastructure with its systems, storage and networks is manifest in IBM's newly introduced "PureSystems" environment. In April, 2012, IBM introduced its PureSystems family of "expert integrated systems". These systems feature immense scale within a single chassis (they can do twice the work of traditional packaged servers by virtue of its very fast internal networking, access to plenty of memory, access to a very large store of data that can be placed within the chassis, and can run multiple hybrid processors). These systems are different from predecessor pre-packaged servers in that they make use of a concept called "*patterns of expertise*" — expert design/configuration knowledge that is used to simplify the deployment and management of a workload, and to improve performance. (These patterns make it possible to configure and deploy systems in about 4 hours as opposed to days or weeks). With PureSystems, IT buyers can achieve very high performance from the outset (without having to figure out how to tune these systems themselves). PureSystems design is a perfect example of how a vendor builds an "integrated infrastructure" and integrates compute nodes with networking and a storage subsystem in order to achieve dramatic performance improvements.

> ***In April, 2012 we wrote an in-depth report on IBM's PureFlex system (a member of the PureSystems family) entitled "How IBM Is Transforming the Application Software Market with Pre-packaged PureSystems". What is important to note about IBM's PureFlex system is that it is an IaaS (Infrastructure-as-a-Service) environment — which means that IBM has highly-integrated its infrastructure products with the underlying system/storage/network environment — streamlining application deployment while improving application performance.***

Another Element to Consider: Look Closely at Your Database Design

So far in this report we have shown how business analytics place different demands on information systems (including systems, storage, and networks). We have also described why designing an integrated infrastructure at the outset is vitally important to the long term

success of business analytics deployments. But we also note that it is also important to tune the database for performance.

As we looked more closely at the IBM deployments described earlier, we found that IBM's major advantage over other information technology suppliers is that IBM designs its own hardware and software (including the DB2 database) — so the company has the ability to improve performance at both the hardware and software level. IBM attacks data latency problems by optimizing data storage within its DB2 database family (for both distributed systems and its zEnterprise server environment) by using impressive data compression techniques, by buffering data, and by using solid state drive technology and storage caches. Data compression reduces the amount of storage needed — and data buffering reduces the need to perform as many storage subsystem reads/writes. Meanwhile, solid state disks storage caches move data closer to the processor for expeditious processing.

A closer look at IBM's DB2 databases show that IBM's DB2 leverages a data partition capability that helps maximize parallel computing performance. To improve analytic query performance, IBM has built materialized query tables (MQT) and multi-dimensional clustering (MDC) — programs that accelerate table reads/writes and indexing (more specifically, MDC is designed to help OLAP processing through the use of clustering indexes. IBM also applies deep compression methods (taking up less storage space and accelerating reads) not only to tables as other vendors do, but also to indexes, temp space, log records and backups. Using these facilities, IBM's solutions can rapidly execute simple queries typically assigned to a single processing thread — or complex queries, which can be decomposed into operations that can be executed in parallel. Further, IBM offers functionality that can rewrite complex queries automatically to significantly improve performance.

> ***This section shows the kind of extreme database tuning and design optimization that IBM undertakes to build high-performance database environments for its customers. It also shows why database tuning is vital when it comes to business analytics workload processing.***

How IBM Delivers Integrated Systems, Storage, and Infrastructure for Business Analytics

IBM has a very broad and deep portfolio when it comes to supporting business analytics deployments. In this section, we highlight some of the systems and storage design elements that enable IBM to build highly integrated, high performance business analytics environments.

Systems

Earlier in this report we described IBM's new PureSystems "expert integrated server" environment. We described IBM's PureFlex System (a member of the PureSystems family of systems) as an IaaS environment designed to offer integrated infrastructure services to IT buyers as well as independent software vendors. IBM also offers System z (mainframes) that provide superior economics for large business analytics applications that need to support thousands of users (see our report here on System z business analytics processing advantages). IBM PowerSystems are very strong analytics processors — especially for applications that are compute-intensive or parallelized. And IBM's System X line is known for performance and scalability leadership in the x86-based server market.

Storage

Business analytics is all about issuing queries against databases (in some cases — very large databases). Accordingly, storage plays a key role in business analytics system design.

Tiering (where data is placed — and on what type of storage data is placed) is particularly important when it comes to business analytics. To manage data tiering, IBM has introduced an advanced data management environment known as EasyTier that uses algorithms to examine how data is being used (reference patterns are observed) — and once that data hits certain thresholds, it can be assigned to a hot tier (SSD drives where it can be read and written very quickly), or to cold tiers (mechanical drives where reads/writes take place more slowly). Thus, tiering helps keep the most relevant data closest to the CPU where it can be executed most expeditiously.

We also see tiering as important from a cost of operations perspective. In days gone by, tiering in the distributed computing world was a largely manual activity — so operator involvement was necessary. Tiering data was complex, required special skills, was time consuming — and costly. IT organizations, accordingly, resisted doing a lot of tiering (except in the mainframe world where tiering has existed for the past 30 years) — instead just buying more storage and putting all data on the same tier. Major advances in the automation of distributed architecture storage tiers have taken place over the past two years — these advances have served to minimize the amount of operator intervention needed to tier storage. Further, tiering reduces the cost of storage (because not all data is tier one data). For more on IBM's tiering efforts, see IBM EasyTier.

Summary Observations

The key messages in this report have been:

- No single processor performs all workloads optimally. To illustrate this point we showed several examples of hybrid systems that use different processors to execute different parts of a business analytics workload. By using the right processors for the right workloads, exponential increases in performance can be achieved;

Integrated infrastructure is vital to the success of business analytics deployments. Infrastructure is the highway that supports data flow to and from a processor. It needs to be integrated and tuned to deliver the performance needed to meet existing SLAs — and it must be designed to accommodate future growth. Failure to do this up-front, can lead to scalability and performance issues down the road.

Few IT organizations have the deep skills needed to build highly-integrated systems/storage/network/infrastructure/database environments. Some vendors have the ability to deeply integrate all of these components. IBM is one such vendor.

From our perspective, IBM is uniquely positioned to deliver high performance, highly-integrated systems, storage, and related infrastructure to business analytics buyers because:

- IBM offers three server lines (System z, Power Systems, and System x) — and offers a wide range of specialty processors that speed up analytics processing (including FPGAs, zIIP processors, and more). IBM offers its customers more "choices" when it comes to deploying the best system to execute their analytics tasks than any other vendor;

- In storage, IBM builds its own storage products, and offers advanced tiering facilities (such as its EasyTier product), as well as advanced storage management;
- From a database management perspective, IBM offers advanced data management software products that can manage modest databases through very large Big Data databases (InfoSphere Streams, InfoSphere BigInsights, and more);
- IBM's leading database, DB2, is known for its compression advantages and for high performance; and
- IBM offers a broad and deep portfolio of business analytics applications (*IBM is highly committed to the business analytics marketplace — the company has spent $16B on 30 acquisitions over the past six years* — and the company has invested heavily in architecting rich business analytics infrastructure environments including pre-integrated hardware solutions).

IBM can integrate all of these elements — hardware and software and services — to bring extremely high performing business analytics solutions to its customers. Due to this major advantage, we do not see IBM's competitors as being as well positioned as IBM to serve the business analytics marketplace.

As for parting advice, we suggest that before choosing a point or piece-part solution from another vendor, your organization owes it to itself to consider the impact and benefits that can be achieved by deploying business analytics software on an expertly tuned integrated infrastructure.

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-6662

This report was developed by Clabby Analytics with IBM assistance and funding. This report may utilize information, including publicly available data, provided by various companies and sources, including IBM. The opinions are those of the report's author, and do not necessarily represent IBM's position.