CLABBY ANALYTICS

Advisory

The Case for Running Real-Time Transactional Analytics on IBM z Systems

Executive Summary

Most enterprises deploy their IBM z Systems as high-volume transaction servers. But, when it comes to analytics processing, these same enterprises usually transfer their mainframe transactional data to other, distributed server platforms for analysis. This practice is known as the extract, transform and load (ETL) process. It is highly inefficient and extremely costly – and it has got to stop! (For a deeper discussion of the problems with the ETL process, see this report. Note: one example in this report shows how one enterprise was able to save $8 million during a four-year timespan simply by eliminating the need to ETL one terabyte of data per day).

The reason we are recommending that enterprises stop moving their enterprise data from the mainframe to other platforms is because IBM's new mainframes are now better suited than ever before to process in-transaction analytics workloads. With improvements in mainframe compute-intensive performance combined with the ability to conduct analytics on mainframe data in real-time (transactional analytics), the mainframe is uniquely positioned as an integrated heavy input/output (I/O) transaction processor that can conduct business analytics in real time.

The new mainframe (dubbed the "z13") now offers more than three times more memory than the previous generation EC12, significantly more cache, greater parallelism and a faster I/O subsystem. As compared to its predecessor, the new mainframe offers 40% more capacity, more than 300% more memory and 100% greater memory bandwidth to storage.

With the introduction of the new mainframe, there are now several good reasons to process transactional ***and*** analytical workloads on the same system. These include:

- *Significant cost savings* – because additional systems, storage and networking equipment does not need to be purchased for analytics processing – and because mainframe millions-of-instructions-per-second (MIPS) are not wasted transferring data to other platforms;
- *Faster results* – because data no longer needs to be transferred to other platforms, the mainframe can perform near real time analysis on the transactional data that it has already captured;
- *Less complexity* – data can be managed all in one place instead of being duplicated across separated data warehouses/data marts. Further, database managers do not have to while transfer data (helping to overcome latency); transform data (reducing errors); nor does data have to be re-synchronized (ensuring that the data is fresh and accurate).

Each of these reasons on its own justifies ending the ETL practice. But all of these reasons together create a very compelling case to stop the practice of ETLing data to distributed servers.

In this *Advisory*, *Clabby Analytics* describes how IBM's new mainframe environment has changed to make it better suited for processing both transactional ***and*** analytics workloads. We discuss technical/architectural changes to the new z13 that make it better suited for processing analytics workloads; we consider the supporting business analytics ecosystem; and, we discuss the relevance of mainframe benchmarks. After considering all of this information, we hope that the lesson learned by our readers is this: ***it no longer makes sense to move data off of the mainframe for analytics processing on distributed servers*.**

Redefining the Mainframe as an Integrated Transaction and Analytics Engine

Over the past several years, enhancements to the z processor and mainframe system design have transformed z Systems. It has long been an excellent general workload processing environment – but now it is particularly well suited for high-volume transaction processing *as well as* complex analytics processing.

We see five enhancements to the new mainframe that make it better suited than ever before for analytics processing. These enhancements include:

1. Dynamic multi-threading;
2. Large memory pools;
3. Accelerated analytics processing;
4. Extended computational performance; and,
5. Data compression acceleration.

These enhancements – and the benefits that they deliver – are illustrated in Figure 1.

Figure 1 – Enhancements to the z Systems That Make Them Well Suited for Processing Analytics Workloads

Large Memory Pools	Dynamic Multi-threading	Accelerated Analytics Processing	Data Compression Acceleration
Access and analyze large datasets in real time instantly	Boost performance for Linux, Java, and zIIP workloads	Optimization of complex, numerically-intensive analytics queries	Capture new opportunities due to lower cost of keeping data online
Up to 10TB of data to deliver up to 50% reduction in response time	24 to 1 consolidation ratio from x86 to zNext for up to 70% lower TCA	Significant throughput and response time improvement for analytics workloads	Reduce storage cost for sequential data by up to 75%

Source: IBM Corporation – January, 2015

Changes at the Processor Level

For decades, the z processor has been a single threaded, stacking CPU – a design that is particularly effective and efficient when it comes to processing transactions. But now, with processor improvements such as the ability to handle more threads and with single instruction, and with

multiple data set (SIMD) improvements, this processor is now better suited to handle compute-intensive workloads. With new dynamic multi-threading capabilities, the z processor can now execute multiple threads simultaneously (as opposed to its single thread orientation in previous models) – enabling the new z Systems to boost performance when handling Linux, Java and zIIP (z Integrated Information Processor – a specialized Java processing environment) workloads. And, because of this threads/performance boost, the new z Systems are better positioned to handle x86 server consolidation (with a 24 to 1 consolidation ratio).

Also consider that the ability to exploit SIMD (single instruction, multiple data set) vector processing is also important in the new z Systems – enabling z Systems to claim best of breed single thread performance, and best cache to thread ration in the industry.

Changes in System Design – Large Memory Pools

To handle multiple queries, while also processing transactions and other workloads, z Systems needed system design improvements – particularly in the amount of memory supported, the amount of cache available, and in the speed of the input/output (I/O) subsystem that feeds data to memory from storage. z Systems got a huge memory boost when the z12 was introduced four years ago (with up to 3TB of main memory). *The new z Systems, the has more than tripled the amount of memory (to 10TB) – and offers greatly increased cache, and support for even faster I/O speeds*. Further, the speed at which data can be delivered has also increased.

z Systems can now access and analyze large data sets in real-time. With improvements in available memory, cache and I/O speed the new z Systems can legitimately be called a unified, integrated transaction/analytics processing environment.

Accelerated Analytics Processing

Last year we wrote a report that described some of the progress that we were seeing in z System design as it pertained to analytics processing. In that report we stated that "IBM introduced IEEE binary floating point facilities at the end of the 1990s. The early 2000s brought 64-bit computing and superscalar parallelism to the mainframe (superscalar architecture implements a form of parallelism called "instruction level parallelism" ─ allowing a single processor to process work at a rate faster than its clock rate). Also, clock speeds have continually increased. Further, we noted that IBM had added "out-of-order execution," and has substantially improved floating point performance (mainframe floating point now rivals reduced instruction set processors such as POWER, SPARC, and Itanium). And, with the introduction of the EC12, IBM added significantly more on-chip cache. All of these improvements contribute strongly to positioning the IBM mainframe as an excellent processor for compute-intensive [numerically-intensive] SIMD vector processing."

The new z13 builds on all of this activity by optimizing the processing of complex, numerically intensive analytics queries – greatly improving throughput and response times for analytics queries.

Data Compression/Acceleration

The new z Systems also feature strong data compression/acceleration enhancements. With compression features, storage costs for sequential data can be reduced by up to 75%. As for acceleration, IBM offers an acceleration appliance known as IBM's DB2 Analytics Accelerator that

speeds the processing of complex analytics workloads. This tightly coupled accelerator uses an IBM server known as IBM's Pure Data System for Analytics as its base and adds software elements that enable the accelerator to be treated as an extention of DB2 for z/OS. It features the use of field programmable gate arrays (FPGAs) that speeds communications between the mainframe and the accelerator, as well as speeds the processing of data. As for the data, it is snapshotted to the accelerator and then constantly refreshed such that the data that is being analyzed is constantly kept current.

Recent improvements to the IBM DB2 Analytics Accelerator include:

- The ability to accelerate a broader spectrum of queries, including support for Static SQL, multiple-row FETCH and multiple encodings on the same accelerator;
- Improved workload balancing, including incremental update performance and improved monitoring; and,
- Improved storage performance, including built-in restore, and better access control of archived partitions and protection for moved partitions.

Customer Comments on the DB2 Analytics Accelerator

Petrol d.d., a mainframe user, is the principal supplier of oil and other energy products to the Slovenian market – and offers a broad range of automotive goods and services, as well as household items, food products and other merchandise. The company uses its mainframe for transaction processing, but also had a need to analyze its transactional data from its retail stores to improve sales. The company chose to use IBM's DB2 Analytics Accelerator to accelerate its IBM Cognos software query performance, offloading the mainframe from complex query processing and thus speeding-up query performance.

According to Pavel Batista, the company's chief information officer "IBM provides us with tools that align with smarter commerce, enabling us to deliver the right message to the right person at the right time, to understand product affinities and intelligently drive the sale, all in a customer-centric way."

Swiss Mobiliar is a leading insurer in Switzerland. The company offers insurance products through 80 agencies throughout the country. To maximize profitability, the company desired to perform analytics on its mainframe transactional data – but also desired not to increase operational expenses. Tto do this, the company chose to use IBM's DB2 Analytics Accelerator. According to Swiss Mobiliar, IBM's DB2 Analytics Accelerator and its z196 mainframe work together to bring transaction processing and analytics together to create a cost effective analytics solution. As for results, the DB2 Analytics Accelerator enables speedy processing of queries with no increase in active server cores. Further, it accelerates 50 percent of queries issued by a factor of 100X; and reduces transaction response times by 20 percent.

Banca Carige Group (Banca Carige) is one of the largest banks in Italy. Like the other examples above, the company chose to use a mainframe to process large volumes of transactions. But Banca Carige also needed to identify new market opportunities, and then create offering to meet customer demand and to attract new customers. To identify these new opportunities, the company needed to analyze its mainframe transactional Big Data. IBM's DB2 Analytics Accelerator was chosen in order to accelerate the speed at which Bana Carige queries could be executed.

The z System: The Supporting Analytics Software Ecosystem

About five years ago IBM executive management mandated that the company's software products are to be developed to work across all IBM servers (this included the mainframe, the System x, and Power Systems). Previously, IBM's Cognos, SPSS and other analytics products had favored distributed systems – but the company's cross-platform mandate made this no longer the case. So the same reporting tools, utilities and applications that were perfected and honed to perform analytics on distributed systems are now available and have been optimized for use on the mainframe.

Additionally, IBM has built IT analytics software specifically for the zSystem. For example, consider IBM's zAware environment – an application that uses analytics to create a model of mainframe behavior, and then identifies changes that have negatively impacted the mainframe. And IBM recently launched its leading Big Data Solution on z with IBM InfoSphere Big Insights

IBM partners are also getting involved in building out the analytics software on z ecosystem (for instance, see this report that we recently published on Veristorm – a maker of a Hadoop database offering for the mainframe environment).

Down with the ETL Process!

The ETL practice of transferring data from the mainframe to distributed servers began back in the 1980s. At that time the mainframe was not designed to serve as a data mart, a data warehouse nor as a business intelligence server. Instead, the mainframe processor had been designed to rapidly process large volumes of stacked transactional data. Distributed servers, on the other hand, with shared nothing architectures, floating point logic, and other computation-oriented features were better designed to process mathematics-oriented compute-intensive tasks (such as business intelligence and analytics).

> ***Given this situation, it was perfectly logical to transfer data to distributed servers to process analytics and business intelligence workloads. Given improvements in mainframe processor and systems design, however, the practice of extracting, transferring and loading multiple copies of a mainframe database to distributed servers no longer makes sense.***

With all of the improvements described in the previous two sections, it is now possible to efficiently perform transaction processing and analytics simultaneously on the mainframe. As a result, enterprises can capture *real-time insights* (as opposed to having to move data to distributed systems for processing, and the having to wait for results) garnered from their transactional data. With the ability to analyze data in real time, enterprises can improve customer service, establish competitive advantage, discover new insights, reduce risk and more. (For example – by using real-time analytics, fraud can be detected before it happens as opposed to using the pay-and-chase method).

Yet, even with all of these improvements, we believe that many enterprises will continue to ETL data. The reasons why most enterprises won't end their ETL processes are:

1. The ETL practice has become instantiated over decades as the way to deploy and manage data marts/data warehouses – and the way to process business analytics applications. *It is a learned behavior that needs to be unlearned*;
2. Many data centers have a mainframe organization and a distributed systems organization – each defending its own turf (the distributed computing organization fights tooth-and-nail not to concede any processing to the mainframers – and vice versa). This is why *executive management must step in and referee this situation*; and,

3. There is a misperception that it costs more money to process analytics on the mainframe. Mainframe MIPS (millions of instructions per second) are considered more expensive than distributed system MIPs. But, in reality, it can cost millions of dollars to buy additional distributed servers, additional storage, additional communications equipment and related software in order to drive the ETL process. And sending data to external servers isn't free either (transferring data burns mainframe MIPs too).

We believe, however, that enterprise executives who do the math will more than likely find that processing analytics applications on the mainframe is far less expensive from a capital equipment perspective, as well as process perspective than ETLing data to distributed systems. Further, operational complexity – and related costs can also be reduced by analyzing data at the mainframe level.

Some of the reasons an enterprise may wish to reexamine its ETL process include cost, time, risk and inefficiency. ETL is costly because the enterprise must invest in additional equipment and software in order to implement and manage the ETL process. By not sending mainframe data to networked external servers enterprises can reduce the number of servers they have to purchase; they can eliminate the need to invest in related communications equipment and software; they can reduce complexity; they can lower management costs; lower power and cooling costs; and they can dramatically reduce the cost of data movement and make it as transparent and coherent as possible. Further, the ETL process is time consuming because it takes a long time to transfer terabytes of data over a network and it takes time to get analytics results back (as contrasted with real-time transaction/analytics processing by not ETLing). It's risky because data can be intercepted; and it's risky because data can become corrupted when it leaves z control. And, finally, the ETL process is inefficient because to protect against data loss and corruption, database administrators often make several copies of the database (sometimes 5 or more) – and creating so many copies wastes storage space and also creates data management headaches (for instance, with so many copies, which copy represents the single, uncorrupted, unmodified version of the truth?).

One Final Comment: The Benchmarking Situation

To forestall enterprise efforts to eliminate their ETL practices, distributed systems vendors will point to industry standard benchmarks in an attempt to prove that distributed systems offer superior performance when compared to a mainframe. There is, however, a major flaw that occurs in benchmarking – and that flaw puts mainframe architecture at a distinct disadvantage in industry standard benchmark testing. We call this flaw the "shared everything/shared nothing" dilemma.

Here's the situation: mainframes are rarely tested in industry standard benchmark scenarios because these benchmarks tend to favor shared-nothing, distributed systems designs such as those used by Oracle UltraSPARC, IBM Power Systems, and Intel Itanium and x86 systems. These shared nothing designs can take advantage of dedicated resources that can be tuned to the extreme to perform a particular workload with very little "sharing" overhead to deal with. In contrast, the mainframe design is a shared-everything design – meaning that the mainframe has been designed to run a wide range of workloads, and when a workload is completed, the compute resources used by that resource are returned to a common virtualized pool. There is systems management overhead associated with this shared everything design – and this overhead places the mainframe at a major disadvantage when compared to shared nothing designs. As a result, mainframes are not usually benchmarked against the other leading commercial microprocessor/systems architectures.

In benchmarks, distributed servers perform much differently than they do in the real world, because in the real world distributed servers are usually burdened with overhead associated with resource sharing. Our recommendation regarding benchmarking, therefore, is this: use industry standard

benchmarks as a performance "indicator." A better approach, however, to determine the performance of a given system is to find anecdotal performance information based on vendor or independent lab tests as well as real-world customer environments. Run benchmarks in vendor labs (most major vendors have facilities for benchmarking). Or, ultimately, conduct a proof-of-concept (PoC) test at your site.

Summary Observations

The central point of this report has to do with the INTEGRATION of analytics with the business transactions. With the new system, it is now possible to process large amounts of data on the mainframe in a performance envelope that is competitive with traditional distributed servers – without sacrificing service levels in the operational environment. Enterprise data can now be processed in real-time without having to move that data to other systems (and without having to deal with latency and management issues related to the ETL process).

In the end, we believe that the ultimate decision factor on where your organization's data should be analyzed should be based on where that data is located. If the data is on – or in close proximity to the mainframe – and if the mainframe is capable of performing the type of analytics required – that data should be analyzed by the mainframe.

> ***IT executives also need to consider that the type of service level agreements that are emerging require that business analysis takes place in real-time. Many enterprises can no longer tolerate the latency and management problems caused by ETL data movement or calls to distributed systems in order to obtain insights.***

Also, enterprise IT executives need to consider that ETLing data to other platforms to save money may in fact be more costly than paying for a few additional mainframe MIPS. And, moving that data creates process complexity and an opportunity for that data to be compromised, corrupted, or mismanaged. It is not wise to ETL data unless it is absolutely necessary (for instance, if there is a particular analytics application that is required – and that analytics application doesn't run on a mainframe).

ETLing data is a hard habit for enterprises to break – it is a process that has been established for decades; it is a common practice; and it is a bone-of-contention in the battle between organizational factions that are trying to defend their mainframe or distributed computing turf. ETLing is costly, time consuming and risky – and it can be avoided by analyzing data on the system that owns that data. Enlightened IT executives who have done the math – and who know how costly the ETL process really is – have broken their ETL chains and are now achieving real-time analytics results with less cost and less risk.

> ***Is the mainframe well suited to conduct analytics processing? If that question was posed five years ago, Clabby Analytics would have answered "no." But now, with great improvements at both the processor and server level, our answer is a resounding "yes!" With the arrival of IBM's new zSystems, Clabby Analytics can now vouchsafe that the mainframe is well suited for both data- and compute-intensive analytics workloads. It is time to stop ETLing data and instead process that data where it can be executed more quickly with less risk – on the mainframe.***

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