CLABBY ANALYTICS

Advisory

The ETL Problem

Executive Summary

There is a common belief in the information technology (IT) marketplace that the cost to transfer data between disparate systems (extract it, transform it, and load it – or "ETL" it) is inconsequential. After all – the logic goes – it can't cost that much to send a bunch of bits over a network to other systems, can it?

> ***Our answer, however, is "it can indeed cost a bundle of money to constantly move data from one system to another – so if you're moving data, you'd better be able to justify that cost".***

The cost penalties for ETLing data can be found in three areas:

1. The additional servers, storage and networking equipment needed to support file transfers and process data;
2. The labor hours involved in managing file transfers and associated data; and,
3. The costs pertaining to wasted system cycles (system mis-utilization).

How much money are we talking about? Depending on the amount of data being transferred (and duplicated), the additional systems and labor costs can run into the millions of dollars – a huge expense just to move data from one place to another and manage it. In one study (included herein) the cost to transfer and manage data between a mainframe and associated reduced instruction set computing (RISC) systems exceeded $8 million over a four year period.

In a recent briefing, IBM shared some of its own findings on ETL costs with a group of several IT industry research analysts. The point that IBM wanted to make was that, *in many cases, it makes no sense to move data off of a mainframe to other systems for processing*. IBM also wanted to emphasize that the mainframe can process a wide variety of analytics workloads without having to move the data to other systems.

> ***But our key take-away from this briefing was more broadly applicable: enterprises need to take a closer look at their ETL costs regardless of which platform they are moving data from (whether it be a mainframe, a RISC machine, or even an x86 server) – because ETLing data can be a very expensive proposition.***

In this *Advisory*, *Clabby Analytics* shares some of the data that we gathered from IBM's ETL presentation – and we share our own perspective on why we believe that one of the first questions that IT executives should ask when trying to determine where a workload should be placed is: "*where does the data reside*?"

Our Approach to System Selection

Our most fundamental belief when it comes to system selection is that no single micro-processor/server environment does all jobs the most optimally – so we believe in matching workloads to the microprocessor/server environment best suited to serve them. This involves evaluating workloads and associated Quality-of-Service requirements, and then matching the workload demands to the processors/systems designs that can most optimally serve those workload.

But, before we even start with a workload evaluation, we need the answer to one important question: "where's the data located?" The reason this is so important is that we want to avoid having to move data from one platform to another. Our reason: it can cost really big money to move data (extract, transform and load data) to other systems. Further, in many cases, enterprises incur a performance penalty when ETLing data.

> ***So, if your data is located behind a mainframe, the first thing we do is evaluate how a given workload will run on a mainframe. If the data is owned by distributed Power Systems, the first place we look to run that workload is on a Power System. And if the data is behind distributed x86 servers, the first thing that we evaluate is how the workload will perform on x86 systems. We do this to avoid very significant extract, transform and load cost and performance penalties.***

Once we know where the data is, we look more closely at the characteristics of a given workload. We want to know how it will exploit the microprocessor and how it will use system resources.

> ***The primary goal for any information system selected should be to achieve balanced performance between processors, memory, and I/O (input/output). Accordingly, we recommend that IT executives try to avoid scenarios where a workload overpowers the processor, memory or the I/O subsystem. To avoid scenarios like this, we recommend that IT executives look at CPU performance characteristics, off-load (to other types of processors) characteristics, memory utilization, processor core efficiency, execution styles, instruction set characteristics, and communications and network facilities. We also recommend that the system design be scrutinized. Is the system design shared everything or shared nothing? What QoS extensions have been designed into the system (such as redundant components for availability, or autonomic management environments that can predict failures before they occur)? What are the power consumption characteristics? Is the system design purpose-optimized? And so on…***

We firmly believe that enterprises that follow this guidance will be able to "workload optimize" their computing environments – driving the cost to compute down to its absolute minimum. Conversely, enterprises that don't follow this guidance …

About the Data in This Report

For years we've been looking for good quantitative and competitive data that supports our view that data proximity (how close the data is to a server) and data ownership (which server environment owns the data) are extremely important considerations to be weighed when selecting a server. To our delight, IBM's Competitive Project Office undertook its own study; gathered data about the ETL process; measured how long it takes to move and

manage data – and has come up with some very reasonable numbers that show how much it actually costs to transfer data from a system that owns the data to ancillary systems.

> ***Readers have a right to question the source of this data. It could be argued that IBM's Competitive Project Office is supporting its own internal agenda when it argues that mainframes should be at the center of many organizations analytics strategy. For those who take this view, however, consider this: IBM has essentially created a model that shows IT executives how to ascertain the ETL costs within their own organizations. Even if you don't believe the numbers that were presented in IBM's briefing (shown later in this Advisory), you will at least have a model that you can use to tally your own ETL costs.***
>
> ***As for us, we find the numbers reported and the methodology used in IBM's analyst briefing to conservative cost estimates and to be fully credible.***

IBM's Real Agenda: We Want the "Mainframe Quarantine" Lifted

The purpose of IBM's analyst briefing was to enlist the support of leading industry analysts in breaking the "mainframe quarantine problem". You see, IBM's mainframe (System z) group has a chip on its shoulder. This organization builds a system architecture that can operate at 100% utilization for long, sustained periods of time; it offers the strongest security in the commercial server marketplace; it has incredible scalability, reliability, availability and resilience – and yet IT organizations often take data that has been captured by the mainframe and move it to distributed servers for processing. This must drive mainframe systems engineers and developers nuts – and rightfully so – because in many cases the mainframe can process that data more efficiently and more cost effectively than external distributed systems environments.

> ***In fact, IBM's System z organization claims that many enterprise IT executives have "quarantined" mainframe systems. For decades the mainframe has been used as the main computing engine for transaction and batch processing – as well as for run-the-business custom and packaged enterprise applications. Unfortunately, given its strong performance in handling these types of workloads, some IT executives have chosen to pigeon-hole the mainframe as a transaction/batch engine and business/custom application processor – essentially quarantining the mainframe from handling the new generation of business analytics applications. IBM wants to change this thinking…***

IBM estimates that 50-60% of the world's operational data resides under System z mainframe control. Vast amounts of financial, retail, and governmental data is held within mainframe databases. ***From our perspective, this means that most of this data should be analyzed by mainframes***. But, instead, a lot of data analysis is being performed on distributed systems running IBM, SAP, Teradata, Microsoft, Oracle, and Sybase data warehouses and business intelligence software. Why is this so?

Part of the answer can be found as far back at the 1980s when a huge backlog of mainframe applications forced IT and departmental organizations to offload some computing tasks to distributed Unix servers. Sales, manufacturing, financials, human resources, distribution and other key applications made their way onto Unix servers. Further, some of these departments even maintained their own databases (creating separate silos of data within organizations). The distributed computing wave was followed by the client/server wave –

a move to host more and more intelligence on smart PC clients (this had a similar erosive effect on Unix as Unix had had on the mainframe). By the beginning of the 1990s there was even talk that the mainframe's days were numbered – that Unix and Windows distributed environments would ultimately become the preferred computing architectures of the future.

To make matters even tougher for the mainframe, a lot of new application environments (such as data warehousing and business analytics) were being developed by independent software vendors (ISVs) – and they placed these applications first and foremost on non-mainframe architectures. Several data warehouse solutions found their way onto Unix/-Windows platforms – far fewer found their way onto mainframes.

Since the 1980s a whole generation of IT managers has grown up with knowledge of reduced instruction set computing (RISC) architectures – and with knowledge of complex instruction set computing (CISC) architectures (more precisely, x86 architecture). These managers know little about mainframe architecture – and often resist mainframe adoption based on this lack of knowledge. Distributed servers are chosen because these managers believe that distributed servers are comparatively inexpensive (see this report which refutes this idea); they believe that offloading the mainframe will reduce costs (because mainframe millions of instructions per second – MIPS – cost more than distributed MIPS); and they believe that the cost of transferring data is insignificant (this will be disproven later in this report).

What these IT managers need to understand is this: 1) where the data resides should have a huge impact on which system should be chosen to process it; and, 2) IBM's System z now has a very rich portfolio of data warehouse/business analytics software solutions that can deliver more results more quickly than leading distributed computing solutions.

What Has Changed to Make the Mainframe a Stellar Analytics Server?

IBM met the distributed systems and client server challenges of the 1980s and 1990s by making the mainframe more powerful (currently the fastest processors in the industry), by significantly improving its I/O subsystem, by adding vast amounts of main memory (currently at 3TB) – and by introducing special workload processors (these special purpose processors include IBM's System zIIP for fast processing of DB2 workloads, IBM's zAAP ─ for fast processing of Java workloads, and IBM's IFL for Linux consolidation and new Linux workloads). Further, IBM has created turnkey, mainframe-based analytics server environments such as IBM's zEnterprise Analytics System 9700 to reduce the time it takes to deploy and run business analytics applications.

These technical improvements (fast processors, fast and plentiful I/O, and large memory) are exactly what is needed to make the mainframe a strong processing environment for analytics workloads. Specialty processors, the introduction of the Linux operating environment to allow for Linux application capture, and a huge IBM commitment to the development of data warehouse and business analytics applications on the mainframe complete the picture by allowing a rich portfolio of advanced analytics applications to be hosted on an environment that is well suited to support such applications.

With fast processors, fast I/O, and large memory, the mainframe is well positioned from a hardware perspective to execute simple queries on the operational data that it already owns. Plus, IBM's software portfolio includes tools and utilities that can be used to prepare data

for analysis; to perform different types of analysis (for instance, IBM's Cognos environment is excellent at performing ad hoc queries, operational analytics, predefined reporting and on-line analytics processing [OLAP]); to perform predictive analysis (IBM's SPSS); and/or to offer cubing services (IBM's InfoSphere Warehouse). With this powerful systems architecture combined with a broad and deep suite of database and analytics tools, the mainframe is very well positioned to handle a variety of queries types.

But is the mainframe also outstanding at performing advanced and deep analytics workloads? The answer is "yes" when you take into account the IBM DB2 Analytics Accelerator (or Analytics Accelerator), a workload-optimized appliance that builds on IBM's FPGA/Intel-based PureData System for Analytics but is tightly coupled with DB2 for z/OS. This environment enables complex queries of mainframe data – executing these queries in record time with full transparency to applications and data management tooling.

> ***The long and the short of this discussion is that, from a hardware perspective, the mainframe is well designed for querying operational data. Further, with the Analytics Accelerator extension, the mainframe can process complex queries in record time. Add to this the broad portfolio of IBM data warehouse and analytics offerings – and the rapidly growing independent software vendor (ISV) ecosystem that is constantly being expanded on the mainframe – and it is hard to refute that the mainframe is especially well positioned to process analytics workloads.***

The Hidden Gem in IBM's Briefing: The Impact of Relocating Data

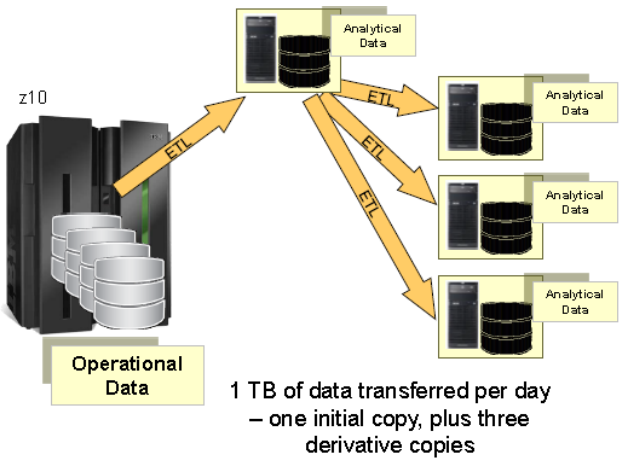
The key lesson to be learned in this report is that the location and ownership of data matters greatly when it comes to reducing computing costs. The problem in many IT shops is that there is a common belief that it costs little to transfer data to other servers for processing. This belief needs closer scrutiny.

What Typically Happens When Transferring Data

Here's a typical data transfer scenario (see Figure 1). In this scenario, the operational data already resides on the mainframe – but this particular enterprise wants to move it off the mainframe to a distributed server for analytics processing. In this case, the data is extracted, transferred and loaded – including one initial copy plus three derivative copies. Immediately apparent should be the amount of additional storage that is required to handle the copy and derivatives transferred. But other costs also enter the picture, including:

1. the mainframe system activity cost for extracting and sending data;
2. the distributed system cost for receiving and loading that data;
3. network equipment and transmission costs; and,
4. system and storage administrative costs.

Figure 1 – A Typical Extract, Transfer, Load Scenario



Source: IBM Corporation, October, 2013

A Closer Look at the ETL Costs

Remember, most IT managers think of ETL costs as insignificant. But IBM's cost model shows a completely different picture. In Figure 2, IBM has modeled systems, storage, transmission and administrative costs for transferring 1 TB of data per day amortized over a four year period. And ***IBM's data shows that ETL costs are VERY significant***!

Figure 2 – The Hidden Costs of ETLing Data

4 yr. amortized cost summary

System z Extract and Send	$2,861,600
Distributed Receive and Load	$4,466,140
Network	$430,408
System z Storage	$49,330
Distributed Storage	$238,720
System z Admin	$22,207
Distributed Admin	$143,090
System z Storage Admin	$5,880
Distributed Storage Admin	$51,960

System costs = $8,046,198

Labor costs = $223,137

Total = $8,269,335

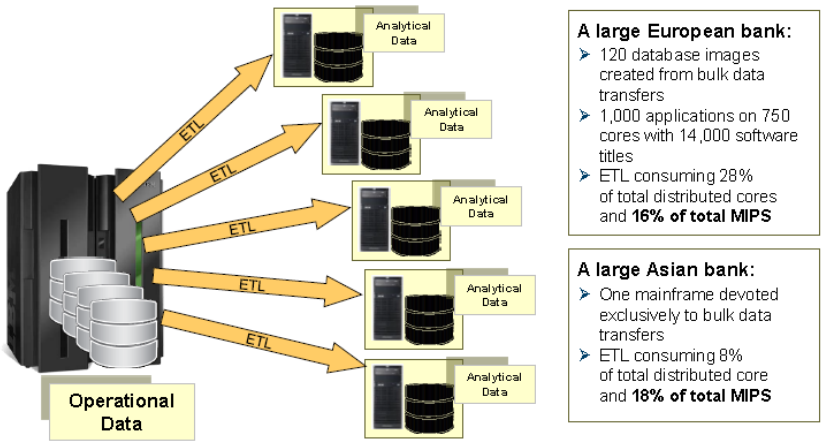
Source: IBM Corporation, October, 2013

Another way to look at the costs involved in ETLing data is to consider the following: assume that all machines in this model are 4 core systems – and that the z10 runs at 85% utilization while the RISC servers run at 60% utilization (these are typical utilization rates). In this case the transfer of data will burn 557 MIPS and use 21 distributed core *per day*. For readers looking to determine their own organizational ETL costs, find out how much each core costs to purchase, operate and support on a daily basis – *and then multiply that cost by 365 days*!

Another Look at ETL Costs from a Customer Perspective

As part of its analyst presentation IBM showed how two customers (a large European and a large Asian bank) handle ETLs to distributed cores. As can be seen in Figure 3, the large European bank creates 120 database images from bulk data transfers – and is running thousands of applications on 750 cores that execute over 14,000 software titles. This bank is expending 28% of its total distributed cores and 16% of its total MIPS ETLing data between a mainframe and distributed environment. *This represents a lot of computing power that is being dedicated to simply relocating and managing data!* The large Asian bank has dedicated an entire mainframe to handling bulk data transfers exclusively – and the ETL activity is consuming 8% of the distributed cores and using only 18% of the total available MIPS. *Again, this represents a lot of computing power that is being dedicated to simply relocating and managing data!* What these examples show is that these banks are wasting a lot of MIPS to simply extract, transform and load data.

Figure 3 – Using a Mainframe to Manage ETL Processes



Source: IBM Corporation, October, 2013

Logically, placing more work on the mainframe where it does not need to be transferred can greatly lower data transfer costs and associated equipment MIPS waste.

Summary Observations

IBM's ETL analyst briefing was designed to show that moving data off the mainframe can be extremely costly – and that if enterprises choose to leave their data on the mainframe, the mainframe can do an excellent job analyzing that data. IBM wanted the analyst audience to know that enterprise IT executives can save a bundle of money if their data is processed by the server environment that owns that data.

One aspect of the IBM briefing that we did not share with you was some of the competetive performance data that IBM put forward on how System z analytics performance compares with other vendor's analytics server implementations (we could not share this data because these test results are IBM company confidential). But we can say that, given IBM's system design, we believe can outperform and out-price/perform its competitors due

to the way that IBM optimizes operational queries on its scalable DB2 z/OS database supported by its industry unique Parallel Sysplex architecture; and, given that IBM offloads deep analytical queries to the balanced and highly parallelized DB2 Analytics Accelerator.

As parting advice, we suggest that IT executives more closely examine the costs associated with moving large amounts of data to mainframes – as well as to other platforms – on a daily basis. Yes, it can be expensive to transfer data from mainframes to other systems for processing – but the same holds true when transferring data from RISC to x86 architectures. In order to reduce computing costs, you should know your ETL costs.

The model for measuring these costs is fairly straightforward. IT executives should measure the daily cost of extracting and sending data; for receiving and loading that data; for network equipment and transmission costs; and, for system and storage administrative costs. Then multiply by these costs by 365 days a year. This model will help your organization approximate its ETL costs. Alternatively, an IBM technical team known as the Eagle team can assist your organization in measuring its ETL costs.

When all ETL costs are finally exposed, expect to be unpleasantly surprised by the final calculation. To fix this situation, place the right workloads on the servers best suited to execute those workloads. And remember to consider that servers that own the data should be given the opportunity to process that data if at all possible in order to avoid expensive ETL costs.

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