The Value of Supplanting ETL with Data Virtualization for the Mainframe

Prepared for Rocket Software by:

David Loshin
Knowledge Integrity, Inc.
March, 2014

Technical Note 2014-02-001

www.knowledge-integrity.com
(301) 754-6350

Introduction

Over the past 25 years, the data warehouse has been institutionalized as a core component of the enterprise information architecture, and in most modern organizations, it would be surprising to not find some form of a data warehouse or a collection of smaller data marts used for queries, reports, and other analytical applications. Yet in the early days of the development of analytical decision support systems (DSS) engineered to meet the need for specialized reports to support strategic decision-making, a practical decision was made to replicate the data used for reporting and analysis to a segregated platform.

Although the original intent of system segregation was to preserve compliance with expected levels of service, the overhead associated with extracting data from the source systems, transforming that data into a form that is suitable for reporting and analysis (also referred to as ETL), and the subsequent loading into the target environment remains the bulk of the effort invested in enabling data warehousing and business intelligence. This suggests a critical question: are the machinations that were devised to address the potential performance deficiencies still required today? In this paper we will explore potential alternatives that can preempt the need for costly effort and infrastructure for ETL while continuing to support expected levels of service.

The decision to segregate the data warehouse from the operational systems hinged on these key drivers:

- **Maintaining response performance**: most mainframe systems were in production supporting transaction processing applications, and there was a concern that the increased demand by reporting and analysis application on the central processing units would degrade the coveted sub-second response time for the transaction systems.
- **Software cost management**: some of the processing for reporting and analysis can be computationally-intensive, leading to the concern of skyrocketing software costs for running the reports on the mainframe.
- **Analytical performance**: the organization of data (in IMS, VSAM, or even relational databases) for transaction processing poses inefficiencies for analytical processing across multiple performance dimensions, including rapid response to ad hoc queries, maintain data consistency between the sources and the targets, timeliness of data delivery and accessibility, and the ability to scale in relation to increased numbers of simultaneous users.

Copying the data from the mainframe to the DSS platforms (as well as their evolved descendants, the data warehouse and the data mart) required a new practice for *extracting* the data from the mainframe, *transforming* the data into a form suitable for reporting and analysis, and *loading* the data into the target environment. This practice, often referred to as ETL, was a capability designed and created to address the need to balance mainframe transactional performance vs. analytical processing.

From the perspective of the business intelligence practitioner, we are left with the impression of the mainframe as a relic completely unsuited to meet the needs of information consumers increasingly

dependent on larger and more varied "big" data analyses. However, architectural enhancements to the mainframe coupled with specialized middleware enable methods for virtualizing mainframe data and making that data directly available to end-user business intelligence tools while avoiding the prohibitive costs associated with utilizing increased mainframe capacity. In this technical note, we consider the challenges introduced as a byproduct of the practice of extracting mainframe data and essentially copying a snapshot of production system data to an analytical platform. We then discuss IBM's System z Integrated Information Processor (zIIP) specialty engine and its practical advantages for reducing on-host integration costs for eligible workloads. We then discuss methods for adapting the mainframe to be used as a platform for serving data to end-user BI tools by enabling it as a virtual appliance for translating external SQL or XML data requests.

Lastly, we consider how enabling data virtualization in a way that directly accesses the mainframe data in place obviates the need for a segregated data warehouse, thereby reducing or even eliminating the need for ETL. In other words, with software written to exploit the zIIP specialty engine, **the cost and complexity of ETL can be eliminated by leaving the mainframe data in place and serving that data directly to the business intelligence front ends.**

Balancing Cost and Complexity for Data Availability

To best understand both the complexity of ETL and its impacts on data availability and utilization, let us consider the typical way that information flows from the mainframe to the data warehouse to address the demand for performance. Data access methods are used to pull the data from the mainframe in its native form, often in flat-file formats augmented with structural indexes (such as VSAM data), with the actual data values often encoded using competing representation standards (such as EBCDIC). The source files are landed onto a temporary staging area platform.

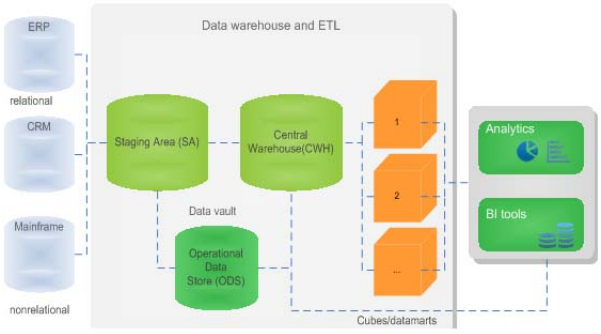

Figure 1: Traditional data integration with data warehouse utilizing ETL to move data

The structures of non-relational data on the mainframe vary. Whether the data is managed using indexed files using the record-oriented VSAM organization, Adabas' inverted list database, or the IMS DB organization with its hierarchical database model, none of these models are particularly suited to efficiently support analytics or business intelligence. That means that any attempt to use mainframe data will typically require a series of data movements as well as a variety of modification in preparation for use, such as storage format transformations (e.g. EBCDIC to ASCII), data edits, standardization, and cleansing are applied. This is followed by reformulation into a relational database model using a denormalized star schema suited to a wide variety of downstream reporting and analysis information consumers.

These processes incur a number of different costs: costs for development of the ETL code, storage and computing resources for the staging area, costs for proprietary software tools supporting the ETL, as well as ongoing execution and maintenance costs. At the same time, additional pressures are showing that in some cases, the ETL approach shows some wear and tear:

- **Data volumes** – There is an increasing interest in analyzing ever-exploding data volumes, but time for moving these data sets from the mainframe are growing beyond Batch window capacity.
- **Data latency** – there is a time lag inherent in the extraction, transformation, and loading process, so data in the reporting system is not synchronized with data in the source system.
- **Consistency** – Because of the data latency, the data in the reporting system will be inconsistent with the data in the transaction platform.
- **Delivery speed** – Shortened time-spans for "return on information" imply that the business needs information faster than Information Technology systems are capable of delivering it.

Overcoming Challenges to Data Availability

Any approach to overcoming the drawbacks of a data integration strategy for mainframe data that relies on extraction and transformation must accomplish two main objectives. First it must satisfy the original motivations for segregation in the first place, namely maintaining response performance for transaction processing applications, not impose increased software costs, and facilitate high performance for the downstream reporting and analytical applications.

Second, it must predictably and reliably address the pain points emerging from the costs and complexity of the conventional ETL approach, such as satisfying requests for massive data volumes without introducing significant delays in delivery and delivering the data in a timely manner. Any solution that effectively addresses these challenges will reduce or even eliminate the inconsistencies and asynchrony that are the byproduct of data extraction.

In retrospect, in many cases the cost and complexity of the ETL process are no longer offset by the assurance of performance for both the mainframe and reporting and analysis application. Yet these issues are not only a byproduct of ETL, they are actually the direct result of introducing ETL in the first place. It may be time to reconsider the wisdom of data segregation so that we no longer have to capitulate to a problem that we created on our own.

Architectural Enhancement with the z Integrated Information Processor

Fortunately, IBM z class systems have architectural enhancements that may eliminate the need for ETL altogether. For IBM System z class mainframes, enhancements originally introduced to open the mainframe to support specialized workloads have matured and have been adapted to different purposes. Examples include the Integrated Facility for Linux and supporting Java execution using the System z Application Assist Processor (zAAP). One of these, the System z Integrated Information

Processor (zIIP), was designed to offload certain types of DB2 data workloads and increase computing capacity while lowering total cost of operations.

According to IBM's documentation, "*the IBM System z Integrated Information Processor (zIIP) is designed to help free up general computing capacity and lower overall total cost of computing for select data and transaction processing workloads for a wide and varied group of eligible workloads including business intelligence (BI), ERP and CRM…*"[1] One can configure an IBM z-class machine with a number of zIIP engines as long as that number is less than or equal to the mainframe's number of central processors (CPs).

There are clear practical advantages of the zIIP specialty engine: processing that runs on the zIIP specialty engine does not count against the mainframe capacity measured in millions of instructions per second (MIPS) a measure of the mainframe's general purpose processor (GPP) performance.That means that processing-intensive tasks such as data transformation, can execute on the zIIP platform, providing three critical benefits:

1. Cost efficiency: because there are little or no MIPS consumed, there is no increase in the system rating, and therefore the overall cost of operations decreases, and
2. Increased capacity: the generally low-cost of a zIIP engine dramatically increases the availability of computing resources, and offloading tasks to the zIIP frees up GPP capacity to perform other critical tasks.[2]
3. Reduced MIPs capacity usage forestalls the need for a mainframe upgrade, which can trigger additional license fees from mainframe software vendors with installed mainframe products.

Data Virtualization on the Mainframe

The key to supplanting ETL as a mainframe data integration strategy is data virtualization. According to noted data virtualization expert Rick van der Lans, "*data virtualization is the technology that offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data stored in a heterogeneous set of data stores*."[3] For the z class IBM mainframes, data virtualization can be enabled by employing the zIIP as a processing engine for transforming data store on the mainframe to directly satisfy the requests for downstream business intelligence and visualization tools. This strategy eliminates the need to move data and significantly reduces mainframe costs.

Mainframe data virtualization implies the need for a data integration tool that combines two important capabilities:

[1] "IBM Systems and Technology Data Sheet for the zEnterprise EC12 (zEC12)," July 2013, Downloaded on Feb 12 2014 from http://public.dhe.ibm.com/common/ssi/ecm/en/zsd03029usen/ZSD03029USEN.PDF
[2] "Adding Some zIIP to Your Mainframe," IBM Systems Magazine November 2007, downloaded on Feb 12 2014 from http://www.ibmsystemsmag.com/mainframe/administrator/performance/add_some_zIIP_to_your_mainframe/
[3] "van der Lans, Rick F., "Data Virtualization for Business Intelligence Systems," 2012 Morgan Kaufmann

- SQL support: This is the ability to translate ANSI SQL-92 compliant queries into requests compatible with a broad range of mainframe data sources – data, programs and 3270 screens (including, but not limited to Adabas, DB2, IMS DB, VSAM, CICS, IMS TM, IDMS, and Natural).
- Low Cost/High Performance: The second involves adapting the use of the zIIP to divert processing costs while running without restriction on its processing speed. Add to this parallel Input/Output data architecture to enable concurrent processing within the z/OS address space. The zIIP engines share the same coherent memory platform with the data as the CPs within the Central Electronic Complex (CEC), and the preponderance of CPU cores within the arrangement of integrated zIIP engines enables scalability in parallel data access to the mainframe files.

The effect is that the mainframe data virtualization on the zIIP provides real-time data integration with true ANSI SQL-92 functions yet *leaves the data in its original location, and reduces mainframe processing overhead* This provides specific advantages for efficiency, performance, consistency, and improved TCO. Network traffic is reduced to communication between the requesting application and the mainframe source, as the need to move data from the mainframe to the series of off-host staging areas and operational data stores (ODS) is eliminated. Redundant data accesses are no longer needed, as the data requests are directly serviced at the mainframe host. Data latency is compressed, since the interaction between data source and processor appliance is a matter of relatively high-speed disk-to-memory or memory-to-memory transfers.

Most importantly, data virtualization on the mainframe lowers the total cost of operations. Application tasks executing on the zIIP engines do not incur the mainframe system software consumption charges. Eliminating the need for costly staging areas or an ODS reduces hardware costs and IT staff resources, and lowers risk by eliminating data replication and data migration.

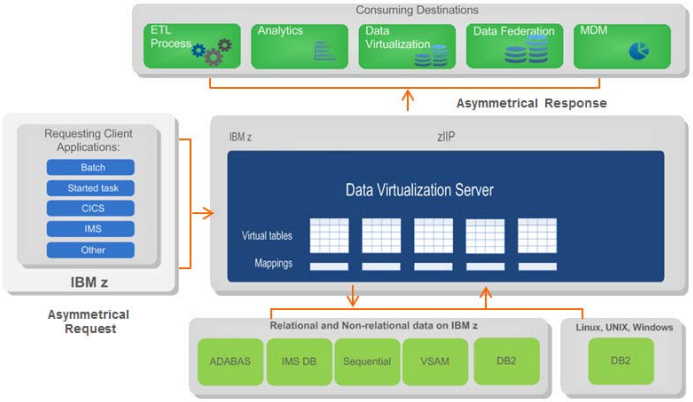

Figure 2: Data Virtualization Server running within Mainframe System z

Reflecting on the value proposition of mainframe data virtualization, we can see that the ability to enable direct access to mainframe data in place obviates the need for continuing to sustain the ETL architecture that is inherently complex and built on outdated data movement scenarios. From the BI practitioner's perspective, data virtualization is an optimized approach to data availability and integration that obviates the requirement for the older-style approaches to extracting, copying, transforming, (perhaps copying again and again), and then loading data into a segregated data warehouse. However, for System z architectures this can only be accomplished using specialized software that transforms the zIIP integrated processor into a mainframe data virtualization appliance. Doing so preserves the query response performance for both transaction processing and analytical

applications while reducing overall mainframe costs to provide seamless real-time interoperability of data for real-time delivery of business intelligence and actionable knowledge.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data quality, master data management, and business intelligence. David is a prolific author regarding best practices for data management, business intelligence, and analytics, and has written numerous books and papers on these topics. Most recently, he is the author of "Big Data Analytics" (Morgan Kaufmann 2013). His book, "Business Intelligence: The Savvy Manager's Guide" (June 2003) has been hailed as a resource allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing, and how all of the pieces work together." He is the author of "Master Data Management," which has been endorsed by data management industry leaders, and the recently-released "The Practitioner's Guide to Data Quality Improvement," focusing on practical processes for improving information utility. Visit http://dataqualitybook.com for more insights on data management.

David can be reached at loshin@knowledge-integrity.com.

About Rocket Data

Rocket Software is a leading global developer of software products that help corporations, government agencies and other organizations reach their technology and business goals. 1,100 Rocketeers on five continents are focused on building and delivering solutions for more than 10,000 customers and partners – and five million end users.

Rocket Software and its Rocket Data products provide real-time data virtualization to enable mainframe relational and non-relational data to seamlessly integrate with Big Data, Analytics, and Web/Mobile initiatives; eliminating the need to move or replicate data, and with significantly reduced costs, complexity and risk.

- Real-time ANSI 92-SQL access to mainframe data for faster, more cost-efficient analytics.
- Industry standard interfaces including ODBC, JDBC, ADO.NET, SOAP, REST, JSON/BSON
- Reduced mainframe TCO -engineered to divert up to 99% of its integration related processing to the System z Integrated Information Processor (zIIP).
- High performance parallel I/O data architecture for continuous data streaming and buffering
- Map/Reduce query optimization

Our customers tell us that IBM System z—the mainframe—is still the best platform in the world for running their critical business applications. And those applications generate and access large data volumes—big data. Increasingly, those applications and data must connect with other applications within the enterprise and even outside the enterprise. Rocket has deep domain expertise and world-class technology to keep the data where it belongs and move the analytics closer to the data.