Enriching Big Data with Mainframe Data Using Data Virtualization

A Whitepaper

Rick F. van der Lans
Independent Business Intelligence Analyst
R20/Consultancy

July 2014

Sponsored by

Rocket

Table of Contents

1 Management Summary 1

2 Big Data in a Nutshell 1

3 Business Benefits of Enriching Big Data with Mainframe Data 3

4 Examples of Enriching Big Data with Mainframe Data 4

5 Integrating Big Data and Mainframe Data by Copying Data 5

6 Data Virtualization Allows for Agile Integration of Big Data and Mainframe Data 6

7 Rocket Data Virtualization Server 7

About the Author Rick F. van der Lans 9

About Rocket Software, Inc. 9

1 Management Summary

Big Data Needs to be Enriched – *Big data systems*, such as click-stream applications, sensor-based applications, and image-processing applications, store amounts of data magnitudes larger than those in more traditional applications. For many organizations the value of these big data systems can be summed up in one word: *analytics*. Big data systems allow organizations to analyze data in ways never been possible before.

New big data storage technology is capable of analyzing massive amounts of data in a fraction of the time needed with more classic technology. And every day new tools are introduced to analyze and investigate all that big data. Unfortunately, most of the big data systems don't contain all the required data needed for analytics. Analyzing big data only rarely ever paints a full picture.

Mainframe Data – More data is needed, and most of it has been gathered by organizations in their classic IT systems, and in many organizations that data is stored on *mainframes*. Some may think that mainframes are dead. But they are not. And what's more important, *mainframe data* is definitely not dead. Conclusion, to fully exploit the value of the investment in big data systems and analytical tools, big data has to be integrated with other data sources, including the ones on the mainframe.

Data Virtualization for Integration Big Data and Mainframe Data – Different technologies exist for integrating multiple data sources. With *data virtualization* big data and mainframe data can be integrated on demand and efficiently. In this case, many forms of reporting and analytics can be supported without the need to copy and store data redundantly, and allowing applications to integrate big data with live mainframe data. Data virtualization servers make it possible to enrich big data with mainframe data.

Accessing mainframe data sources requires integration technology that understands the mainframe and its peculiarities. *Rocket Data Virtualization Server* (Rocket DVS) has been designed and optimized to access mainframe data sources. It's its core strength. For example, because it runs on zIIP processors, Rocket DVS doesn't interfere with mission critical enterprise applications running on the general purpose central processors and thus effectively reduces mainframe processing usage and reduces costs. Rocket DVS makes using the System z platform cost effective for data virtualization.

Rocket DVS doesn't interfere with mission critical enterprise applications running on the general purpose central processors.

2 Big Data in a Nutshell

Big Data and Analytics – In the world of data management, big data and NoSQL databases are increasingly gaining acceptance. According to Information Week's 2014 State of Database Technology survey[1], 13% of those surveyed are running Hadoop. And the MongoDB NoSQL database server demonstrated its market dominance showing up as the only NoSQL database in the top ten of the survey.

[1] Information Week, *2014 State of Database Technology*, March 2014.

Big data systems store amounts of data magnitudes larger than those in more traditional applications. For example, click-stream applications, sensor-based applications, and image-processing applications, all generate massive numbers of records per day. The amount of records stored surpasses more often than not hundreds of millions of records.

The business value of most of these big data systems can be summed up in one word: *analytics*. Due to the new types of data and the level of detail of the stored data, big data systems introduce new forms of analytics. For example, cities can use live camera feeds to manage traffic flow automatically, social media data can be analyzed to minimize future customer complaints, and weblogs can be studied to avoid order cancellations.

Examples of Big Data Systems

Example 1 - Weblog Records: Successful websites generate massive amounts of weblog records. Every time a visitor clicks on a button or moves to another webpage, a weblog record is written to a file or database. These weblog records can be used afterwards to analyze usage of the website: do visitors pick the shortest route, how long do they view a page, and so on. In addition, these records can be used instantly to influence what's been shown on the next page. For example, if they view pages with certain products, maybe an advertisement of these same products should be included on the next page they open.

Example 2 - Customer Call Transcripts: Customers communicating with call centers usually reveal significant amounts of information about their views, sentiments, likes, and dislikes. They also give opinions on products and services. Especially when all these calls are transcripted, they form a valuable source for analytics. By applying analytics, customer experience can be improved, more revenues from more effective cross-sell and up-sell efforts can be generated, and real-time feedback to customers can be provided. Analytics of all this big data can also help to improve agent performance. By analyzing call characteristics, the call transcripts, the strengths and weaknesses of specific agents can be determined.

Example 3 - Social media data: Blogs, tweets, Facebook messages and informational Websites, a wealth of information is hidden in the vast amounts of data being created every day on social media platforms and the Internet. Unfortunately, many organizations today have barely touched the surface of exploiting this data for analysis. As Spangler and Keulen[2] succinctly described in their book *Mining the Talk*, organizations are not listening:

> *"People are talking about your business every day. Are you listening?*
> *Your customers are talking. They're talking about you to your face and behind your back. They're saying how much they like you, and how much they hate you. They're describing what they wish you would do for them, and what the competition is already doing for them. They are writing emails to you, posting blogs about you, and discussing you endlessly in public forums.*
> *Are you listening?"*

Example 4 - Sensor Data: Machines and devices are more and more equipped with sensors. These sensors are turning them into *smart* machines. All this automatically generated data can be used to analyze and optimize processes. For example, data from machine-based sensors can be used to fix problems before

[2] S. Spangler and J. Keulen, *Mining the Talk, Unlocking the Business Value in Unstructured Information*, IBM Press, 2008.

they occur, GPS-based sensor data can be deployed to optimize truck routes, and camera data can be used to manage the traffic flow in a city by influencing traffic lights. Smart machines generate more data which enriches an organization's analytical capabilities. The *Internet of Things* will even accelerate and amplify this trend.

Big Data Technology – The sheer amount of big data has a direct impact on the database technology used. For this reason, organizations have invested in data storage technologies other than the familiar and traditional SQL database servers. Many of them selected *Hadoop* and *NoSQL* technology. All these powerful technologies have many advantages, but they share one disadvantage: they don't support the database lingua franca called *SQL*, making them difficult to access using standard reporting and analytical tools.

3 Business Benefits of Enriching Big Data with Mainframe Data

Is Big Data "Big" Enough? – Big data storage technology is capable of analyzing massive amounts of data in a fraction of the time needed with more classic technology. And every day, new tools are introduced to analyze and investigate all that big data. So far, so good. Unfortunately, most of the big data systems don't contain all the required data needed for analytics. The missing data makes it impossible to show the full picture. For example, a big data system containing weblog records may show how customers navigate a website, but it can't tell how loyal this customer is. That type of data is stored in the CRM system. Or, a big data system may contain all the tweeted complaints of particular customers, but it doesn't know what the sales figures are for those same customers, nor how many ordered products they returned. Another example relates to sensor data, which is usually cryptic. For users to understand sensor data, it has to be enriched with descriptive data.

To get a full picture of customers, sales, and website-usage, to really understand sensor or weblog data, big data is usually not "big" enough. More data is needed, and most of it has been gathered by organizations in their classic IT systems, and in many organizations that data is stored on mainframes.

Is the Mainframe Dead? – Every few years, one can read the mythical phrase "the mainframe is dead." When client/server was introduced in the 1990s, this phrase became quite popular[3]. Again, when the Internet was introduced, the demise of the mainframe was announced. More recently, with the Cloud the future of the mainframe was forecasted as cloudy[4].

However, we must conclude that, although mainframes look different from 20+ years ago, they're not dead. It is true that IBM's market of proprietary hardware recently hit a bump. At the beginning of 2014, IBM's revenue's for System z mainframes[5] were down with 37%. Nevertheless, it's still a multi-billion business. Meanwhile, according to a 2011 survey among 1,347 large mainframe users[6], 93% of the global

[3] S. Lohr, *Why Old Technologies Are Still Kicking*, March 2008, The New York Times, see http://www.nytimes.com/2008/03/23/technology/23digi.html?_r=0

[4] D. Norfolk, *Stop Press: Mainframe Finally Dead…*, July 2011, see http://www.bloorresearch.com/blog/the-norfolk-punt/2011/7/stop-press-mainframe-finally-dead/

[5] See http://searchdatacenter.techtarget.com/news/2240213002/IBM-cloud-efforts-intensify-as-System-z-Power-hardware-revenues-tank

[6] Computer Weekly, *Mainframe Spending to Rise, Survey Finds*, September 2011, see http://www.computerweekly.com/news/2240105592/Mainframe-spending-to-rise-survey-finds

IT executives cited the mainframe as a robust, long-term solution in their enterprise IT strategy.

But what's more important, and this is regardless of the current commercial success of the mainframe market, most applications running on mainframes can't be migrated to another platform without a major and costly redesign and rewrite.

The Priceless Value of Mainframe Data – So, mainframes are here to stay, they're not dead. And what's more important, the data they manage is definitely not dead.

As long as mainframes have existed, organizations have invested heavily in the development of IT systems for them. This is especially true for financial organizations. For example, systems for servicing loans, processing cash deposits and withdrawals, and processing payments and cheques, form the backbone of retail banks. They collect the most essential data a bank can own. Their dependency on mainframes is so high, that if all the mainframes on this planet would magically disappear overnight, the entire financial world would grind to a halt within hours.

Countless organizations still manage their most critical enterprise data with mainframe-based IT systems. Without these systems, organizations would be lost. Despite all the hype around big data and the introduction of new data storage technologies and platforms, enterprise data stored and managed on mainframes is still crucial and priceless.

4 Examples of Enriching Big Data with Mainframe Data

The same examples described in Section 2 are used here to explain the business value of enriching big data with mainframe data.

Example 1 – Weblog Records – Weblog records show how visitors navigate a website. They show the products that visitors are interested in. How long did the visitors study particular products? Did they navigate the website efficiently? At what point did they cancel their purchase order? All this information is extremely valuable to an organization. But weblog records don't show how much this customer has bought in the brick-and-mortar stores, so sales via the website and sales in the stores can't be easily combined. The weblog records don't show how many complaints the customers have sent by email, or how many products they returned. All that additional data may be needed to determine how an organization should react to the customer's behavior. It can also determine the urgency with which issues should be resolved. A large portion of this data is stored on the mainframe.

Example 2 – Customer Call Transcripts – Analyzing customer call transcripts gives insight in the intention of their calls. Was the intention purely inquiry, was it a complaint, was it to issue an operational instruction, or was it an intent to purchase a product or service? In addition, the quality of service can be monitored by analyzing these transcripts. Analyzing the transcripts may also show, for example, whether the organization is accused of something, or whether there is an intention by the customer to churn.
What applies to analyzing weblog records, applies to customer call transcripts as well. To make the right decisions, a comprehensive picture of a customer is required. In other words, an enhanced 360 degrees view of a customer is mandatory. To show such a 360 degrees view, mainframe data is essential.

Example 3 Social Media Data – Analyzing remarks and complaints made by customers on social media platforms can be insightful. But again, without a full picture of these customers, the result may be misleading. For example, relationships between customers may be crucial to understand certain remarks. However, these relationships may not be known in the big data system containing all the social media messages. A CRM system running on the mainframe may be aware of these relationships. Another example is when organizations want to react differently to customers based on whether they are good or bad customers. But to make that qualification, data managed by the mainframe-based sales system must be accessed.

Example 4 Sensor Data – Most of the sensor data coming from machines is highly cryptic. Specific codes indicate machines, locations, and sensors. Even measurement data may be coded. The descriptive data needed to understand this cryptic data may reside in another system. To be able to analyze it, sensor and descriptive data must be combined. The same applies to a GPS-based tracking system. That type of data indicates the location of a truck or device, but it doesn't tell who the driver is and what the planned route is. Nor does the GPS-data indicate that the truck is off-route. The planned route must be known for more complex forms of analysis, and that type of data may be stored in a mainframe-based routing system.

Summary – In many situations, big data by itself is not sufficient for analytics. It doesn't paint a full picture. In many cases, data from other systems is needed to enrich the big data. And many organizations have that data stored in and managed by more classic IT systems running on mainframes. The challenge is to seamlessly integrate big data stored in new systems with data sources available on the mainframe.

5 Integrating Big Data and Mainframe Data by Copying Data

The need to integrate big data with mainframe data is obvious. As indicated, analyzing only big data rarely ever paints a full picture. To get that full picture, big data must be integrated with other data sources. And in numerous organizations those data sources are stored on the mainframe. Thus, the technological challenge is how to integrate them with big data systems.

Copying data using ETL tools is the solution that comes to mind first. With an ETL tool mainframe data can be copied to a big data source for analytical purposes. This approach will technically work, but it means that the analytical tools integrate big data with slightly outdated mainframe data. In many big data systems, especially the ones dealing with weblog data, sensor data, and social media data, time is of the essence. In fact, microseconds can mean the difference between success and failure. Therefore, copying data may not be the right option, because with this approach the analytical tools don't access 100% up-to-date mainframe data, but outdated copies.

An alternative approach is to copy the big data to the mainframe and do the integration there. This is not a desired approach, because copying big data may take long and may lead to high mainframe storage costs.

6 Data Virtualization Allows for Agile Integration of Big Data and Mainframe Data

Data Virtualization in a Nutshell – An alternative approach to integrate big data and mainframe data is through *data virtualization*. Data virtualization is a technology for on-demand integration, transformation, and manipulation of data that's available in all kinds of data sources and to present all that data as one logical database. When data virtualization is applied, an abstraction and encapsulation layer is provided that, for applications, hides most of the technical aspects of how and where data is stored; see Figure 1. Because of this layer, applications don't need to know where all the data is physically stored, whether it's on a mainframe or not, how the data should be integrated, where the database servers run, how to insert and update the data, what the required APIs are, which database language to use, and so on. When data virtualization is deployed, to every application it feels as if one large database is accessed.

Data Virtualization and Big Data – Data sources accessible by data virtualization servers are not limited to SQL database servers. Big data stored in Hadoop and NoSQL data storage systems as well as data in mainframe database servers can be accessed. All these data sources can be presented to applications as one logical database. The data virtualization server handles all the conversions and transformations of the concepts and languages supported by these data sources to, for example, SQL. All reporting and analytical tools can easily work with *all* the data and it makes analytics of big data enriched with mainframe data possible.

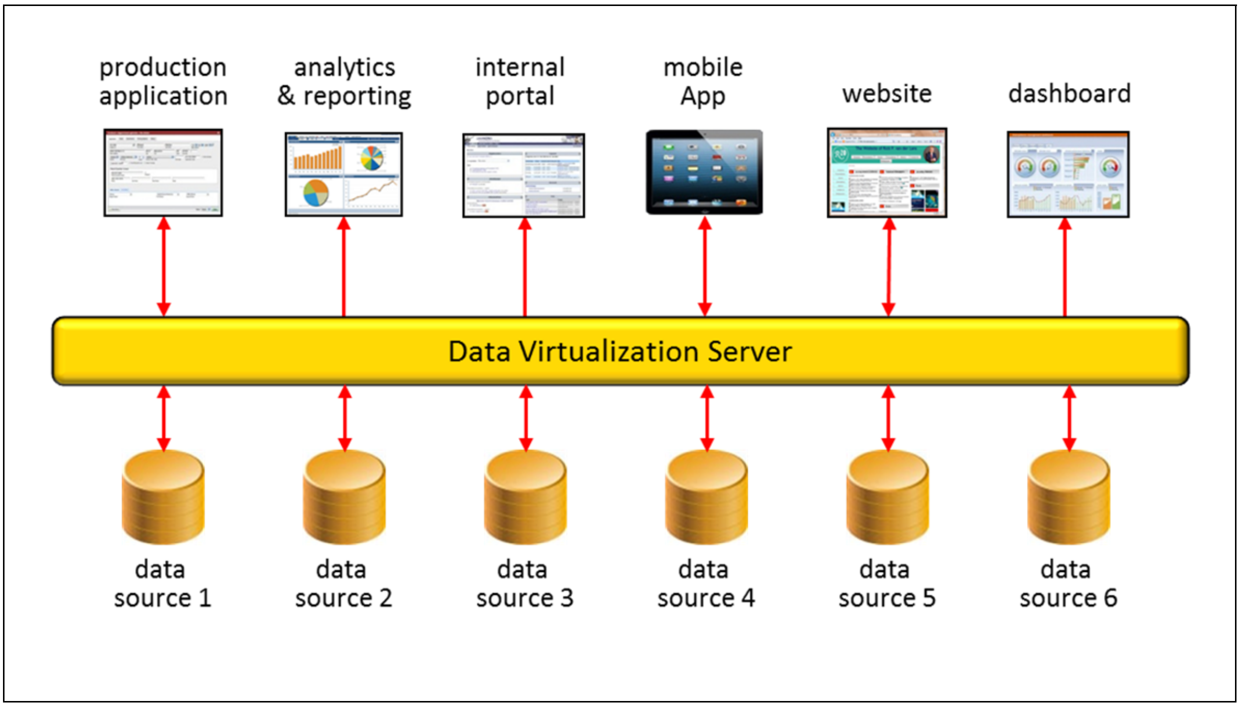

Figure 1 *Data virtualization servers make a heterogeneous set of data sources look like one logical database to the applications. The data virtualization server is responsible for data integration.*

Because data virtualization servers use on-demand integration, the data available for analytics is operational data, not outdated data. In addition, because data doesn't have to be copied and stored redundantly before it can be analyzed, the solution is lightweight and therefore more agile. If analytical requirements change, the integration solution developed with a data virtualization server can be changed rapidly.

7 Rocket Data Virtualization Server

Rocket DVS – *Rocket Data Virtualization Server* (Rocket DVS) is a data virtualization server designed specifically to integrate data sources on the mainframe data and to integrate mainframe data with off-mainframe data. Rocket DVS can efficiently access all the well-known mainframe database servers, such as CA IDMS, IBM CICS, DB2 z/OS, IMS TM & DB, Software AG Adabas and Natural. In addition, typical mainframe file systems, such as sequential files and VSAM files, can be accessed. The product has taken advantage of IBM's data integration standard DRDA (Distributed Relational Database Architecture) by which it can access many SQL database servers, including IBM DB2 and PureData (Netezza), Oracle, and Microsoft SQL Server.

To streamline integration with big data, Rocket DVS includes a capability for MongoDB's NoSQL database that transforms mainframe data into a binary form of JavaScript Object Notation (JSON) referred to as BSON. MongoDB uses JSON documents in order to store records, similar to how tables and rows store records in SQL database. MongoDB represents these JSON documents in a binary-encoded format called BSON[7]. BSON extends the JSON model to provide additional data types and to be efficient for encoding and decoding within different languages.

The Architecture of Rocket DVS – To efficiently access the mainframe data sources, Rocket DVS runs on the *zIIP processor*[8] (System Z Integrated Information Processor). These processors are designed to handle specialized workloads, such as large queries on DB2, Java, and Linux, and divert processing away from the mainframe's central processors. The importance is that by running on zIIP processors, Rocket DVS doesn't interfere with mission critical enterprise applications running on the general purpose central processors and thus effectively reduces mainframe processing usage and reduces costs.

The strong points of Rocket DVS are:

- Efficient database access to a wide range of popular mainframe data sources.
- Minimal interference on the mainframe data sources due to generation of highly efficient mainframe code.
- Integration of the data sources takes place on the mainframe itself.
- Existing data security systems are not bypassed.
- Ability to seamlessly integrate mainframe data with MongoDB.

[7] MongoDB Architecture, see http://www.mongodb.com/json-and-bson

[8] See http://www.ibmsystemsmag.com/mainframe/administrator/performance/add_some_zIIP_to_your_mainframe/

Summary – Based on its internal characteristics and architecture, Rocket DVS makes it possible to seamlessly integrate big data stored on typical big data platforms with data stored on the mainframe. Because Rocket DVS uses on-demand integration, applications see live data and not outdated data. In addition, there is no need to copy and store data redundantly. Rocket DVS implements an agile form of integration that makes it easy to change the integration solution when new forms of analytics so dictate.

With Rocket DVS big data can be enriched easily and efficiently with mainframe data, so that a full picture of business objects can be presented. This improves the analytical capabilities of an organization.

Note: For a more detailed description of data virtualization and the Rocket DVS product, we refer to the whitepaper *Making Mainframe Data Available to the Entire Organization with Data Virtualization*.

About the Author Rick F. van der Lans

Rick F. van der Lans is an independent analyst, consultant, author, and lecturer specializing in data warehousing, business intelligence, database technology, and data virtualization. He works for R20/Consultancy (www.r20.nl), a consultancy company he founded in 1987.

Rick is chairman of the annual European Enterprise Data and Business Intelligence Conference (organized annually in London). He writes for the eminent B-eye-Network.com[9] and other websites. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles[10] all published at BeyeNetwork.com. The Data Delivery Platform is an architecture based on data virtualization.

He has written several books on SQL. Published in 1987, his popular *Introduction to SQL*[11] was the first English book on the market devoted entirely to SQL. After more than twenty years, this book is still being sold, and has been translated in several languages, including Chinese, German, and Italian. His latest book[12] *Data Virtualization for Business Intelligence Systems* was published in 2012.

For more information please visit www.r20.nl, or email to rick@r20.nl. You can also get in touch with him via LinkedIn and via Twitter @Rick_vanderlans.

About Rocket Software, Inc.

Rocket Software is a global software development firm that builds enterprise products and delivers enterprise solutions in the following segments: Business Information and Analytics; Storage, Networks, and Compliance; Application Development, Integration, and Modernization; and Database Servers and Tools. Rocket is engaged in business and technology partnerships with IBM, EMC, Fujitsu, HP Enterprise Services, Hitachi Data Systems, Avaya, Epicor, and many others. The company is headquartered in Waltham, Massachusetts, USA. For more information[13], visit www.rocketsoftware.com or follow them on Twitter @rocket.

[9] See http://www.b-eye-network.com/channels/5087/articles/

[10] See http://www.b-eye-network.com/channels/5087/view/12495

[11] R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.

[12] R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.