IDC
Analyze the Future

WHITE PAPER

The Mainframe as a Key Platform for Big Data and Analytics

Sponsored by: IBM

Carl W. Olofson Dan Vesset

August 2013

IDC OPINION

As the evidence of the competitive edge enjoyed by organizations that rely on data-driven decision making mounts, big data and analytics (BDA) has become a top agenda item for a growing number of executives. At the same time, hype about big data technology capabilities and inflated promises of outcomes abound. These ignore real challenges faced by organizations of all sizes. Many organizations don't have the competency or maturity to address the range of technology, staffing, process, and data requirements involved in managing big data. With the opportunity to unlock the value of big data to accelerate innovation, drive optimization, and improve compliance comes the need to demonstrate value, re-create business processes, ensure the availability of appropriately skilled staff, and navigate expanding technology alternatives.

When the subject of BDA comes up, most people do not think of the mainframe, yet it has built-in characteristics needed to deploy BDA technology in a well-managed environment — one that embodies many of the requirements of a cloud-ready system. System z has over 40 years of history in evolving virtualization, dynamic resource assignment, and multi-tenancy, all with security characteristics unmatched by distributed systems. In addition, System z supports a wide range of operating systems to increase infrastructure flexibility, including Linux, as well as all the leading programming languages, including Java. It permits rapid re-assignment of physical assets to virtual machines, with sophisticated monitoring and tracking. All this is based on a heritage of technology that has been proven in decades of use of mission-critical applications at some of the largest enterprises in the world. Those who question the suitability of System z for BDA should consider the following:

- The essential architecture of a mainframe system, with its ready-made network of specialized devices, centrally managed and organized, delivers the performance and scalability required for BDA workloads.
- The reliability of the mainframe at a level that distributed systems still cannot match, a result of decades of development and refinement, makes it the ideal platform for mission-critical workloads.
- In most cases, at least some of the transactional data needed for BDA is already being managed and maintained on the mainframe, and the ability of the system to facilitate data sharing between application spaces without impacting performance on the transactional source application results in ready access to up-to-the-minute business data for analysis at the point of action.

These facts support the idea that BDA technologies, such as Hadoop, should find a ready home on System z.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

IN THIS WHITE PAPER

This white paper highlights the importance of assessing big data and analytics (BDA) requirements not as a single, homogeneous requirement but as a range of potential use cases. The paper outlines the requirements for a big data implementation and shows how businesses can successfully address the needs of BDA, enabling organizations to leverage their existing investments in mainframe applications as well as highly skilled existing IT staff to extract insights from a variety of sources and integrate them seamlessly into existing transactional applications on a centralized environment, thereby realizing real-time analytic capabilities.

THE BIG DATA CONUNDRUM

IDC defines BDA as a mix of data, talent, technology, processes, and services that allow for effective management of potentially large volumes of multi-structured and/or high-velocity data. It enables a range of business intelligence and analytic applications to support tactical, operational, and strategic decision-making processes across the organization, delivering greater business value.

BDA solutions continue to demonstrate direct, quantifiable business value documented by a growing number of research studies, as well as by organizations' internal, unpublicized efforts.

For example:

- A study by IDC found that organizations classified as "Fact Finders" — described as more analytically oriented — are 20% more likely to be among leaders within their industry.[1]
- A study by MIT's Center for Digital Business found that organizations that utilize data-driven decision making are 5% more productive and 6% more profitable than their competitors.[2]

Another IDC study (the June 2013 IDC and Computerworld Business Analytics and Big Data Survey) shows that 88% of organizations that have widely deployed analytics and business intelligence have recognized tangible benefits from these projects. For 90% of these organizations, the benefits met or exceeded expectations, and for 82%, the time to achieve quantified benefits met expectations or was shorter than expected.[3]

It is important to recognize that BDA does not represent a single, homogeneous workload and that no single technology can address all BDA requirements.

While much of the BDA market attention is placed on capturing online consumer behavioral data in the work of semi-structured clickstream logs, a wide range of data sources and types contribute to big data. In addition, use cases range from experimentation and ad hoc discovery in laboratory-like environments to those that are operationally mission critical and therefore have specific requirements beyond simply being able to process large amounts of data. These requirements include enterprise-grade reliability, availability, scalability, and security, among others. In other words, big data can include structured, unstructured, and semi-structured data;

 ©2013 IDC

data that is processed in batches at specific intervals and streaming data; and data that could be 100 terabytes or multiple petabytes.

Most enterprises that begin to engage big data technology to solve business problems start small, with a pilot project or two. As the value of Hadoop or other big data technology becomes known, they start to stand up implementations. Each is designed to solve a specific problem and serve a specific group within the enterprise. As these implementations become large, persistent, and needful of ongoing administration, they are handed over to IT, where they can be managed properly. But then another issue becomes critical — scalable resource manageability. Put simply, an escalation in the number of big data environments, such as Hadoop clusters, creates tremendous strain on an already budget-constrained IT staff.

With most enterprises expecting IT to do more with less, and as applications become more complex, databases grow, and the number of analytic databases increases, IT staff are already stretched to their limits. Now, on top of all that, comes this new workload. For example, 30% of organizations indicate that one of the top BDA challenges is managing technology sprawl, and another 30% (in response to a separate question) indicate that a shortage of IT skills for providing the needed hardware infrastructure hinders the success of BDA initiatives.[3]

To understand how big data workloads can pile up and yield an unmanageable quagmire, consider the following: Multiple Hadoop or NoSQL database implementations on clusters of commodity hardware may offer not only savings to individual groups over classic data warehouse technology but also greater flexibility. Some such deployments solve analytic problems that the data warehouse is simply not designed to handle. But as they add up, they become an administrative nightmare for IT, and those that are not in constant use create a mounting inventory of massively underutilized hardware.

Big data deployments involving technology like Hadoop tend to grow data endlessly, yet are not used constantly. If they are deployed in discrete, separately managed clusters, however, those resources are tied up constantly, whether they are in use or not. Even commodity hardware costs money to keep running, and node failure is an ongoing fact of life with such clusters. Also, these systems, developed separately, each on its own cluster, share neither data nor resources, representing still greater inefficiency.

The Challenge of Managing Cloud-Based Big Data

One way of dealing with these issues is to create a private cloud environment where system and storage resources can be virtualized and shared across projects and where services, such as databases, can be spun up and down as needed. Such an approach requires detailed planning and large-scale deployment of carefully designed servers and storage, provisioned with hypervisor and resource management software and managed with coordinated administration. Many IT organizations are not really up to such a challenge in terms of either expertise or staff size. For example, a third of organizations indicate that they lack sufficiently skilled IT staff for BDA projects.[3]

In addition to all this, access to enterprise production data, especially mainframe data, is arm's length, requiring importation and transformation with the associated

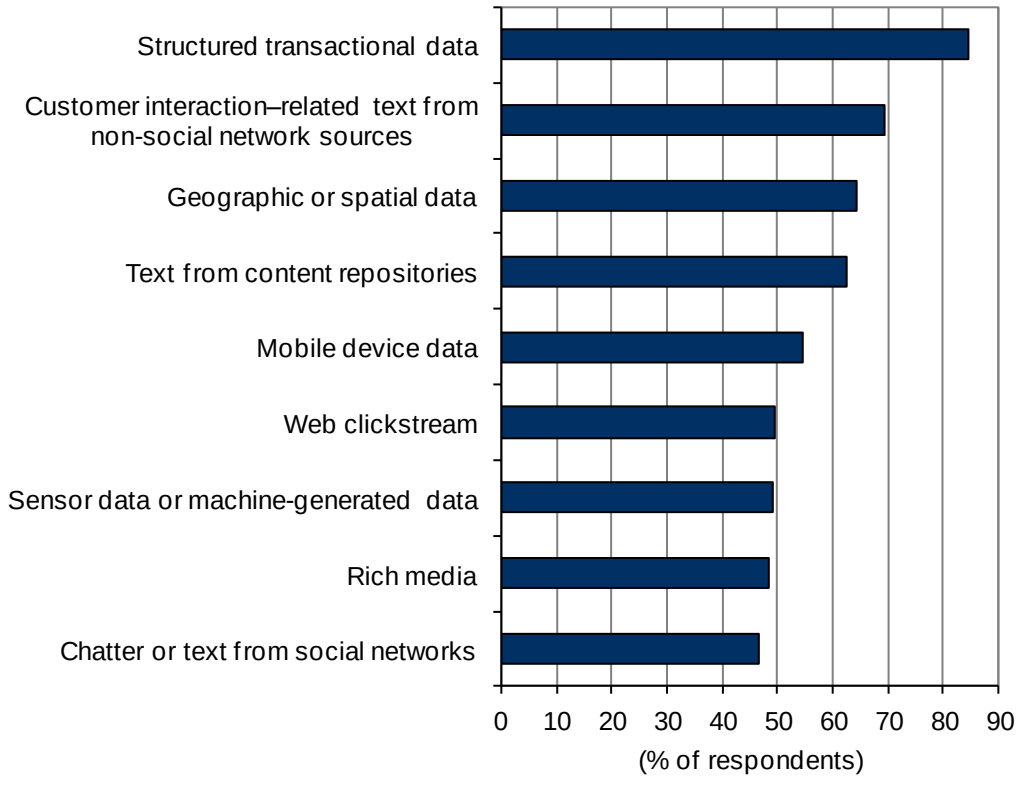
administration of those processes. Data transformation from mainframe to distributed systems is always complicated by the format differences inherent in those platforms (e.g., characters converted from EBCDIC to ASCII, translation of packed decimal numbers, conversion of floating-point numbers). All these operational details add to the overall complexity of management and administration of the processes involved.

Another challenge involves broader governance, integration, and security issues when dealing with big data. Most big data environments, such as Hadoop, offer very little by way of protecting sensitive data or enforcing data management policies. Issues involving such policies, including integration, quality, security, and governance, were all in the top 7 biggest challenges identified by organizations.[3] For example, data in Hadoop cannot be secured by field or type because security is at the file or system level. It would be better to protect the data field by field, controlling access and providing data masking support. Even better would be to run Hadoop in an environment that has ironclad access control and identity management. Additionally, many big data use cases now include data of multiple types from a variety of sources (see Figure 1) — making the previously known data integration challenges even more germane.

FIGURE 1

Data Types Analyzed

Q. Are the following types of data analyzed in your organization?



n = 330

Source: IDC and Computerworld Business Analytics and Big Data Survey, June 2013

 ©2013 IDC

Hadoop and Other Big Data Technologies

Most of the buzz regarding big data has been about Hadoop, but Hadoop is batch oriented and so is not suitable for immediate analysis of streaming data. Also, as a program execution environment, it has no built-in data management features; these are left as exercises to the developers. Therefore, for data collections that require consistency and order, Hadoop can act as an ingestion point but would be a very poor manager. When we consider the different kinds of data to be analyzed and the varied needs for both depth of analysis and speed of delivery of that analysis, it should be clear that Hadoop is simply not the answer to all big data needs.

Hadoop remains useful for large-scale ingestion of unorganized data and for some bulk aggregation analytics. For regular, repeated analysis of structured big data, a natural synergy exists between Hadoop as an ingestion point and the classic RDBMS-based data warehouse as an analytics platform. Other big data technologies that would seem appropriate for various big data workloads include the following:

- Graph databases are better for pattern and relationship analysis; they are faster and more efficient than Hadoop and more flexible than conventional RDBMS.
- So-called "document-oriented" or "object-based" DBMSs, handling self-describing objects such as JSON documents, are interesting for managing data where there is just one application codeset involved and the data organization is highly variable.
- "NewSQL" databases can do both transaction and standard analytic processing and feature elastic scalability, but with the flexibility of changeable schemas.
- There may be other databases, probably in the "NoSQL" category, that will address very demanding analytic workloads not well served by the others.

Needed: Manageable, Reliable, Ready-to-Use Cloud Resources

Big data technologies call for resources and services that can be extremely large at times, but are seldom constantly required, or that have capacity requirements that vary over time. Ideally, one would want the ability to dynamically assign and adjust such services and resources in order to avoid costly over-provisioning. One way of achieving such dynamism is to acquire ready-made cloud configurations or systems and software designed to enable rapid deployment of private clouds. Such products still require setup and management and usually involve technology from multiple vendors, including server, storage, and network hardware providers.

Another approach would be to use a system of resources that are factory built and configured with purpose-built management software, all from one vendor.

SYSTEM Z AS A CLOUD SYSTEM FOR BIG DATA

For those already using System z as a platform for OLTP and analytic applications, managing their data on some combination of IMS and DB2, there is another answer, and it involves resources already at hand. The mainframe can be thought of as a cloud-ready system and in fact has been one since before the cloud concept was commonly adopted.

Virtualization Baked In

The mainframe system virtualizes its physical resources as part of its native operation. Physically, the mainframe is not a single computer but a network of computing components including a central processor with main memory, with channels that manage networks of storage and peripheral devices. The operating system uses symbolic names to enable users to dynamically deploy and redeploy virtual machines, disk volumes, and other resources, making the shared use of common physical resources among many projects a straightforward proposition. Multiple such systems may be blended together in a Sysplex environment.

The mainframe can support Linux virtual machines and Java applications and can arrange them into virtual clusters, providing an environment that could enable Hadoop or various NoSQL technologies to run without modification — depending upon the application. Instead of investing in new and different hardware and software combinations, one can deploy on the existing mainframe system and, if necessary, expand its capacity to deal with the new workloads without incurring a big datacenter architectural change. The IT staff can apply existing expertise to the setup and management of such workloads.

Moreover, it features the capability of moving data, either dynamically or on a scheduled basis, between these environments and existing mainframe applications or data warehouse databases in a system-supported manner. Data from application systems can thus be shared without impacting the performance of IMS, DB2, or CICS processes.

IBM DB2 11 for z/OS, currently in an early support program, provides support for integration with IBM InfoSphere BigInsights (IBM's implementation of Hadoop). IBM Cognos or DB2 for z/OS can invoke predefined HDFS JAQL queries. Even though these might run elsewhere in the enterprise, DB2 for z/OS can retrieve the result sets from the InfoSphere BigInsights environment and load them into DB2 for z/OS tables for integration with the data warehouse and further analysis using the IBM business analytics solutions mentioned previously. There are many similar use cases for IMS. In addition, users could also leverage the Machine Data Accelerator component of BigInsights to analyze IMS system logs and use the results to tune the IMS system, enhancing performance and system health.

Analytics Ready

In the mainframe environment, users can integrate data held in Hadoop with various NoSQL, DB2, and IMS databases in a common environment and analyze that data using analytic mainframe software such as IBM Cognos and SPSS, ILOG, and IBM InfoSphere Warehouse. Mainframe users can take advantage of such factory-integrated capabilities as the following:

- The IBM DB2 Analytics Accelerator for z/OS, which is based on Netezza technology, substantially accelerates queries by transparently offloading certain queries to the massively parallel architecture of the Accelerator appliance. The DB2 for z/OS code recognizes the Accelerator is installed and automatically routes queries that would benefit from this architecture to the appliance. No application changes are required.

- IBM PureData System for Hadoop is a purpose-built, standards-based system that architecturally integrates IBM InfoSphere BigInsights Hadoop-based software, server, and storage into a single system.

- IBM zEnterprise Analytics System (ISAS) 9700/9710 is a mainframe-based, high-performance, integrated software and hardware platform with broad business analytics capabilities to support data warehousing, query, reporting, multi-dimensional analysis, and data and text mining.

- The ability to integrate Real Time Analytics Transactional Scoring in DB2 for z/OS allows for efficient scoring of predictive models within the milliseconds of a transaction by integrating the IBM SPSS Modeler Scoring within IBM DB2 for z/OS.

Flexible Scalability

Unlike mainframe systems of the past that had fixed resource inventories that were difficult to expand, today's IBM mainframe systems such as the IBM zEnterprise EC12 (zEC12) provide configuration options that support both vertical and horizontal scalability, including a variety of different types of processors such as general-purpose Central Processors (CPs), Integrated Facility for Linux (IFLs), System z Application Assist Processors (zAAPs), System z Integrated Information Processors (zIIPs), and Internal Coupling Facilities (ICFs).

The mainframe also can be used as a management and integration platform for IBM Power and System x systems deployed on IBM BladeCenter technology using the zEnterprise BladeCenter Extension (zBX), so BDA deployed in those environments can be integrated with mainframe capabilities under a unified management umbrella.

IBM System z and the zEnterprise Analytics Hub as an OLTP System and Analytics Platform

Another growing requirement seen in the market today involves combining analytics and transaction processing workloads. This hybrid requirement enables organizations to minimize data movement from the OLTP database to an OLAP database, as well as when historical data needs to be integrated with current or real-time data. Often systems with these types of requirements support customer-facing processes ranging from fraud detection to customer services. Although it is not the only use case that can benefit from a mixed workload system, according to IDC research, less than 10% of organizations indicate that the decision support needs of customer-facing employees and operational employees are met to the fullest extent needed by the existing BDA system(s).

By combining IBM DB2 for z/OS with two or more analytic capabilities highlighted previously, IBM is effectively providing a system for mixed OLTP and OLAP workloads — what IBM refers to as online transactional and analytics processing (OLTAP).

FUTURE OUTLOOK

Custom-built platforms designed to support private cloud deployments for big data use cases, some from single vendors and others from combinations of vendors, will arise over the course of the next few years to compete with the System z in this regard. IDC expects this to become a market within the next five years. While many of these platforms will be quite robust, one must consider, if a mainframe is already in the datacenter, whether it makes more sense to take a chance on some new configuration or to deepen one's investment in what is known and has proven reliable for many years — a platform already familiar to the IT staff.

CHALLENGES AND OPPORTUNITIES

There is ongoing pressure in the executive suites of many enterprises to reduce the commitment to mainframe technology because of its perceived cost and dependence on an increasingly scarce talent pool. IBM will be challenged to demonstrate the value of System z (by means of both the arguments in this white paper and other supporting assertions) and to ensure that in the future the pool of talent qualified to manage such systems will increase while making the System z increasingly simple to manage, thereby requiring less specialized knowledge. If IBM can achieve these things, the opportunity for System z as a BDA platform is substantial.

CONCLUSION

While there is no one answer that is right for everyone when it comes to system configurations for big data management, it seems clear that for mainframe users, a powerful solution to at least part of this challenge is readily at hand. Mainframes have the expandability, adaptability, and reliability needed to address

 ©2013 IDC

many of the challenges of big data applications. Users considering big data challenges and how to address them should think about the following:

- Although not all big data projects involve mainframe data, many do; in such cases, a mainframe deployment may make the most sense.
- Deploying on a mainframe means that resources can be accessed for a rich variety of mainframe-based data access and analysis systems.
- Although the mainframe is the original virtualized system environment, it has evolved in ways that fit well with emerging models of data management, including BDA, and organizations can apply it to those new workloads using the skills that IT already uses to manage the rest of the mainframe environment.

REFERENCES

1. *Analytical Orientation and Competitiveness: The Difference Between Fact Finders and Fumblers,* IDC #223408
2. *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?,* MIT Center for Digital Business
3. IDC and Computerworld Business Analytics and Big Data Survey, June 2013

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2013 IDC. Reproduction without written permission is completely forbidden.