JESKELL Systems, LLC | Article

HOW TO REV UP YOUR DATA ANALYTICS PROGRAMS

Three decades on since the advent of the x86 architecture, the computing world has steadily marched toward standard hardware regardless of the computing workload. Even supercomputing algorithms increasingly run on machines made of thousands of x86 modules. This trend followed — and decisively ended — the era of proprietary hardware architectures, operating systems, and programming interfaces.

OR DID IT?

Growing evidence shows that in the burgeoning era of big data and — more crucially — big data analytics, purpose-built systems can yield better results, yet come in at cost levels competitive with the use of commodity x86 hardware.

Advanced data analytics, such as the examination of network data for cybersecurity, is enhanced when the infrastructure is designed, built and optimized for this purpose. Commodity hardware and database tools are unlikely to provide the speed and efficiency required by federal agencies. This white paper will describe a set of software tools that, coupled with specially tuned hardware — supported by the OpenPOWER foundation, so they are not proprietary — can produce power analysis results at competitive lifecycle costs.

Think about the cybersecurity problem. In the past two years, large organizations have been hit with a series of highly publicized cyber attacks. Retailers, health care networks, and federal agencies such as the Office of Personnel Management (OPM) have lost control of millions of individuals' personally identifiable information. Or they've suffered the loss of valuable intellectual property.

In some cases, the breaches have been the result of malware or phishing. In others, they've resulted from malicious players obtaining administrative credentials. Whatever the vulnerability exploited, the results have been the same: loss of data, damage to reputation, and substantial expenses to fix systems and provide credit monitoring and other services to those affected.

Federal agencies operate their IT systems under a layered set of policies. They require continuous monitoring and diagnostics, and increasingly, measurement to detect insider threats. Plus, like all organizations, agencies face a rising tide of sophisticated phishing attacks. Their increasing use of virtualization, agile cloud services, and mobility all expand the so-called surface that is exploitable.

The result? Cybersecurity has become a big data exercise as activity from multiple networks is combined with unstructured data from documents of all types.

Standard relational databases used for transactional applications are not suited to hold the multiplicity of data types an agency needs to analyze to meet cybersecurity requirements. Indeed, the "big" data challenge is less a problem of size and more one of data diversity. Data warehouses may solve the diversity problem when loaded with data run through a format and schema engine for a specific business purpose. However, if these warehouses are built mainly for business trends analysis, they may not be architected to provide real-time alerts or responses to queries necessary to satisfy cybersecurity demands.

Another challenge besides choice of the proper DBMS itself is that, regardless of how fast its read-write cycles may be, a database represents data at rest — a condition that must certainly be part of any cybersecurity data analytics effort. After all, several of the most notorious data breaches have involved wholesale purloining of data at rest.

But a DBMS doesn't allow IT or security operations center staff to complete the other component, namely analysis of data in motion. Comprehensive awareness of live network activities, for example, can enable prevention of the type of spoofed-identity logon that could result in data exfiltration. Or, data in motion analysis could detect suspicious activity — information or documents headed out on your network to unauthorized destinations or by unauthorized users — giving you an early warning of a breach or violation of policy (intentional or non-intentional).

"IBM and Jeskell offer Hadoop bundled with a variety of services, including a cloud-hosted deployment, SQL tools for querying data in your Hadoop file system, analytic tools for when you don't know the specific answer a user is looking for, and visualization tools for making sense of what's in your data."

Beyond the database and data streaming issues lies the legacy data center and its capability for big data analytics. Quite simply, even the most up-to-date data center built out with standard or commodity components may not be up to the dual tasks of analyzing data at rest and data in motion.

True, IT can adopt instances of the open source Hadoop for processing large, diverse data sets and run it on dozens, hundreds, or thousands of commodity machines. But the building of so-called data lakes can introduce additional problems. Power consumption, for one, rises fast and can grow beyond the capacity range of the data center. Large clusters of machines can become difficult to manage. Performance may suffer as a result of the compounding of small latencies. Even when you can pack eight terabytes of data onto a single drive, you've solved a cost challenge but not necessarily an I/O performance one.

That's why a growing number of federal agencies are trying a fresh approach to the infrastructure they require for mission-critical data analytics. The approach links to existing infrastructure and data sources, but provides a new platform optimized for high-speed analysis without disrupting operations of — or investments in — existing data centers. Read on to discover the elements of an analytics infrastructure.

BUILDING BLOCKS OF YOUR ANALYTICS INFRASTRUCTURE

As we've established, big data analytics has become a necessary component of ederal cybersecurity and other critical operations. Working as a major partner to IBM in the federal market, Jeskell has been testing and deploying purpose-built analytical systems using high performance software and hardware combined with our own value-added expertise.

If the purpose of hardware is to run software, the first elements of an analytics infrastructure consist of the software stack — specifically software designed for large and diverse data sets.

Hadoop provides a foundation block for the system we're describing. Hadoop is an open source platform designed for large scale processing on multiple servers. Hadoop is associated with the two operative qualities here — large and diverse. Hadoop handles up to petabytes of information, and can also combine both structured and unstructured data types. The Apache Foundation, home to Hadoop, notes that the distributed file system at the heart of the project is designed for high availability on large clusters of hardware because it senses failures at the application layer running on them all. So a single hardware failure doesn't stop the whole system.

IBM and Jeskell offer Hadoop bundled with a variety of services, including a cloud-hosted deployment, SQL tools for querying data in your Hadoop file system, analytic tools for when you don't know the specific answer a user is looking for, and visualization tools for making sense of what's in your data.

With the underlying file system provided by Hadoop, your analytics infrastructure incorporates other key tools provided under an extensive suite of products known as InfoSphere. Among these tools:

- Streams is essential to analysis of high volume data in motion, for example, geospatial data, network activity, or text such as documents or e-mail. Increasingly, agencies need to analyze sensor-generated data such as from surveillance cameras. Streams takes user-developed applications built under the Java-oriented Eclipse environment, and applies them to data stream sources. In fact, the newest version of Streams, 4.1, lets developers create data streams analytic applications directly in Java. And it is tightly coupled with Apache Spark, an open source big data processing engine.
- SPSS is IBM's suite of predictive analytics applications. It enables organizations to make data-driven decisions for a variety of functions, including hiring, fraud and risk mitigation, and customer interactions. In the cybersecurity domain, SPSS can analyze diverse and large sets of data, looking for clues indicating threats. For example, SPSS can help alert

you to users whose data usage or network activity suddenly changes. The suite includes tools for statistical analysis, an analytic server optimized for Hadoop distributions, and for increasing collaboration on analytic projects.

- Cognos Analytics is IBM's core business intelligence suite. It lets individual users generate insights from data, presenting them in graphic form on desktop or mobile devices. Cognos is available in server and cloud editions, the latter having several options depending on the number of users to whom the agency needs to assign analytic tasks.

A word on databases: As noted, a standard RDBMS used for transactional systems is not likely up to acting as the repository for multiple data sources and types required in analytic situations. The not-only-Sequel Query Language, or NoSQL, database model has emerged as the solution to large scale data analytics. These "flat" models have existed for many years, but have re-emerged in the Web 2.0 era and the advent of big data. Open source products such as Accumulo, MongoDB, and Redis have capitalized on capabilities in the Hadoop/MapReduce complex to let users not only house data, but access large amounts of it using clustered hardware. Products provided by Cloudera and Splunk's Hunk are also underpinned by Hadoop.

Jeskell's CyberSentinel is a turnkey solution built from IBM Research models and the above-mentioned software. CyberSentinel can extend an agency's existing capabilities in security and information management and intrusion detection and prevention. It employs machine learning instead of traditional rule-and-signature models for real-time detection and disruptions of cybersecurity events such as advanced persistent threats — something often missed by other tools.

CyberSentinel brings big data analysis to the cybersecurity challenge, incorporating a number of tools for modeling, machine-based learning, streaming data analysis, and interoperability with many big data repositories. What's the real value to agencies? There's no need to embark on a large-scale custom software integration project. Jeskell's done the work using its own experience in cybersecurity and analytics, and its successful partnership with IBM.

STEP INTO THE DATA LAKE

In the 1990s and until today, CIOs sought to build data marts — repositories of cleansed and structured data for specific analytical purposes. But the data mart concept doesn't scale to where it could support analytics for unlimited purposes such as predictive modeling and trends discovery.

Here again, Web 2.0 companies initiated the idea of the data lake — a large and constantly growing repository for holding multiple sources of data in their original formats. Hadoop supports this view of data. The Internet of Things and other streaming sources build up

fast, along with unstructured data from organization e-mail and documents, plus standard business data from structure systems built around relational databases.

Data lakes differ from data marts and data warehouses in several respects. You normally would process data in some way, say with structure or a schema before it is loaded into the mart. Data goes into a lake as it is, there for sampling or pulling into data analytics applications. In the latter case, it may stay unstructured until it is read by that application.

By implication, data marts tend to get expensive as they grow, whereas the lake concept is optimized for low storage costs.

For cybersecurity purposes, an agency might consider a lake made of low-cost storage and a Hadoop framework, but limit data sources to those relevant to cyber. Cybersecurity analytics requires application of tools to both streaming data and to large batches of data in a typical Hadoop instance, so a lake may incorporate more sources than initially imagined. Note that the addition of new data from streams to the lake changes the nature of what the agency is likely to learn in future analysis.

HARDWARE FOR DATA ANALYSIS

An irony of the Hadoop and big data movements is that, while they were designed for use on low-cost, commodity hardware, that very hardware can sometimes work against the performance requirements of analytics in the service of decision making. Especially so in cybersecurity, when even minutes from knowledge to action can make a difference in whether a breach is successful.

Hadoop is often cited for its ability to store, and via its MapReduce component, provide access to very large datasets, all on commodity x86 software.

This is fine as far as it goes, when you only consider the price of the hardware. But two problems emerge. First, the power consumption of many, perhaps thousands of machines required for a data lake. At a time when federal agencies are under a mandate to continually reduce power consumption per data center and consolidate data centers, it may be hard to justify a large cluster of machines

The second problem centers on disk drives. As CPU power continues to rise, disk access becomes a bottleneck — even as individual disk capacities of eight terabytes are becoming common. Meaning performance gains in classic large x86 clusters are beginning to even out.

Indeed, one major independent IT research firm states flatly that organizations running open source environments, typically Linux, and needing to run increasing workloads must consider alternatives to x86 server farms. Requirements for scalability, agility,

and analytics (along with transaction environments) require more manageable, robust, powerful and performance-oriented hardware infrastructure.

Simply put, standard environments struggle to stand up to the performance and scalability requirements of enterprise analytics applications — in reality, any big data application.

IBM has engineered high-performance systems specifically to enable Hadoop to run faster than what's possible on commodity PC clusters, yet with far easier manageability and lower total power consumption. These systems come in a wide array of capabilities, from workgroup level to enterprise. What they have in common is a version of IBM's proprietary Power8 processor core.

The Power8 chip architecture is a reduced instruction set design capable of massive multithreading — 96 threads on a 12-core microprocessor. Combined with memory caches both on chip and peripheral, the chip is capable of two to three times the speeds of its predecessor, Power7.

Power8-based IBM servers run from single chip to rack-mount machines configurable with up to 80 Power8 cores.

Key to system performance is the Power8 communication capabilities and the storage systems IBM has engineered around them. Imagine a 20-, 40-, or 80-core machine accessing eight terabytes of memory. Moving a ring further out, PowerServers incorporate IBM's CAPI-coherent accelerator processor interface. The port connects to one of many application accelerators on the market. But much more than application accelerators, CAPI's power lies in the way it treats storage and other peripherals on a direct access basis, as if they're memory, rather than over a long, bottlenecking communications channel.

Beyond the processing hardware, big data analytics requires high performance storage. Here, too, a revolution is occurring, driven by falling costs of solid state, or flash memory, drives.

Analytic applications don't like latency, and neither do observers in networked security operations centers. Yet no agency has unlimited money for storage. IBM's big data storage solutions incorporate high performance subsystems with flash at the top of the hierarchy. More importantly, software tools such as IBM Spectrum Storage optimize the elements in a hybrid storage system that includes cloud resources. IBM Spectrum Scale creates software-defined storage specifically to manage unstructured data in the Hadoop Distributed File System with storage tiering based on policy. The result is minimal latency balanced with efficient use of the storage hierarchy — flash down to tape. In some instances it can reduce storage costs by 90 percent.

Jeskell has developed mature expertise in deploying these combined systems — BigInsight and PowerServer plus IBM innovative storage — in federal agencies for a variety of analytic solutions.

DEVELOPING ANALYTIC APPLICATIONS

With an infrastructure built and optimized for analytics, using big data for cybersecurity — or any domain — becomes a matter of applications. Along with their Hadoop distributions, a growing number of companies offer integrated analytic engines that save programming and enhance Hadoop performance on fast hardware. One example is the IBM Data Engine for Analytics tuned for optimal performance with InfoSphere BigInsights and InfoSphere Streams.

> "An irony of the Hadoop and big data movements is that, while they were designed for use on low-cost, commodity hardware, that very hardware can sometimes work against the performance requirements of analytics in the service of decision making. Especially so in cybersecurity, when even minutes from knowledge to action can make a difference in whether a breach is successful."

Open source environments such as Rational Developer for Power Systems Software and the Eclipse-based integrated development environment offer up-to-date pathways to analytics application. They are useful for agile methodologies — something to which a growing number of federal agencies are trying to move. That is, they're adopting development styles that engage end-users throughout the process, while delivering small but frequent modules. What better way to develop analytic applications than by making sure the analysts themselves are instrumental in guiding the way.

In planning a cybersecurity analytics program, it's useful to keep certain technology fundamentals in mind.

1. Analytics is more than business intelligence. It starts with a data structure-agnostic approach, meaning the data can support a wide range of queries from a variety of users that the IT shop may not be aware of at the start of a project. For example, security operations staff might be your initial set of users. They would want to see patterns in network activity. But human resources might have an interest in who is working at what hours. The records management staff might query the data for clues to document usage and flow.

2. The CIO staff should consider the software infrastructure requirements of big data-based analysis, and whether it has the in-house expertise in modern analytics tools. The market is full of Hadoop experts, but they might not be the same people who maintain legacy COBOL systems.
3. If your agency chooses a cloud provider rather than building an in-house analytic infrastructure, or if you go hybrid, make sure your provider has the platforms we've described. Especially in hybrid situations, you want matching capabilities to maintain the agility often attributed to the cloud model.

Growing numbers of companies and federal agencies whose missions are inseparable from their IT systems — like financial services and defense — are moving to the open source/data lake/accelerated hardware model. It's something all agencies need to consider. And Jeskell offers the technical expertise, federal experience, and proven capabilities to be a trusted partner every step of the way.

For More Information:

Jeskell: http://www.jeskell.com/analytics

IBM InfoSphere: https://www.ibm.com/analytics/unified-governance-integration

IBM Streams: http://www-03.ibm.com/software/products/en/ibm-streams

Power8 servers: http://www-03.ibm.com/systems/power/hardware/

OpenPOWER foundation: http://openpowerfoundation.org/