DATAGUISE
SECURE BUSINESS EXECUTION

In partnership with:
Microsoft Azure

SAFELY UNLOCK THE BENEFITS OF BIG DATA IN THE CLOUD

With Microsoft Azure HDInsight and Dataguise DgSecure

DON'T LET CLOUD SECURITY CONCERNS HOLD BACK YOUR BIG DATA INITIATIVES

There's no question that big data and cloud computing have moved beyond the hype and into mainstream adoption.

The popular open source framework, Apache Hadoop, has fundamentally reduced the complexity of how organizations process, store, and analyze huge volumes of structured and unstructured data. Moreover, the economic benefits of storing and processing large data repositories in the cloud have proven to be increasingly compelling over the last few years. Finally, more organizations are now realizing value from their big data projects. In a recent industry survey, 72% of respondents said their big data initiatives exceeded expectations, and nearly 75% said that their businesses would benefit if they could exploit all of their data.[1]

Still, many organizations hesitate to embark on cloud-based big data projects. They wrestle with how to store data in the cloud so that it's accessible, available, and most importantly, secure. They express concerns that demanding big data workloads have unpredictable scale and dispersed components, and must process huge amounts of unorganized data, some of which is confidential and highly sensitive.

This eBook explains why the Microsoft Azure platform and its cloud-based Hadoop service, HDInsight, help solve the challenges associated with storing, processing, and analyzing big data. It also describes how the Dataguise DgSecure solution can deliver precise sensitive data security for the HDInsight service. Together, Dataguise and Microsoft enable organizations to securely leverage highly-available, scalable Hadoop deployments, and safely unlock the business benefits of big data.

[1]Computing Technology Industry Association, Big Data Insights and Opportunities, November 2015.

MICROSOFT AZURE HDINSIGHT: POWERING BIG DATA IN THE CLOUD

Big data is messy. Within a data lake repository one might find petabytes of unstructured files: website clickstreams, satellite images, sensor data, GPS signals, server logs, and other data types. Hadoop is popular precisely because it handles just about any data format. Before Hadoop, processing large datasets required expensive computer hardware. Because Hadoop enables scalable, distributed computing on industry-standard hardware, big data analytics is now within the reach of an increasing number of organizations.

Deploying Hadoop clusters in the Azure cloud is cost effective for other reasons. With Azure HDInsight it's possible to spin up a Hadoop cluster in minutes and add only as many nodes as needed. Because compute and storage costs are based on usage, it's possible to save on operating expenses by provisioning a Hadoop cluster, analyzing data, and then shutting it down when finished.Elastic scaling and pay-as-you-go billing are compelling reasons for using cloud-based Hadoop. What differentiates Azure HDInsight is its integration with other Microsoft Azure services, making it a comprehensive, all-in-one big data solution for many business scenarios.

Consider these capabilities available in the Azure cloud:

- **Leverage business analytics tools.** HDInsight integrates directly with Microsoft business intelligence tools, such as Power BI, and Excel, enabling users to employ sophisticated analysis and visualization capabilities to Hadoop data.
- **Connect on-premises and cloud-based Hadoop clusters.** By using the Hortonworks Data Platform on Azure it's possible to transfer Hadoop data between an on-premises datacenter and the Azure cloud. With the Microsoft Analytics Platform System, organizations can query on-premises and cloud-based Hadoop clusters simultaneously.Process real-time events. HDInsight includes a full complement of Hadoop components, including Apache Storm, an open-source stream analytics tool for processing a millions of events in real time.
- **Apply advanced analytics capabilities.** Combine the Microsoft advanced machine learning technologies of Cortana Analytics Services with the HDInsight data lake service to build prognostic modeling tools for scenarios such as energy demand forecasting and predictive maintenance.
- **Run HDInsight in Azure on Linux or Windows clusters.** Organizations that already use Hadoop on-premises on Linux can use Linux tools and templates to extend their deployment to Azure in a hybrid cloud scenario.

HDInsight enables faster access to healthcare information

"If you think about each clinician that struggles with getting timely, accurate data, and you compound it on a national scale, then it becomes an immense challenge. You have lots of small pieces of data coming in from multiple places, and it can be very difficult to aggregate and interpret. The processing power you would need to handle all of that information is beyond the capability of most organizations. A hospital can't just stand up a server farm to process millions of case notes from an emergency care system in addition to other data.

With a solution based on SQtL Server 2012 and Windows Azure HDInsight Service, we can capture data written in plain English and use it to improve services, instead of waiting for a resource-hungry process to collect and code the information. Our healthcare surveillance system based on SQL Server 2012 provides actionable insight—it's about compressing the time it takes to identify the problem and act on it the best way possible."

—Paul Henderson, Business Intelligence Division Head, Ascribe Ltd.

DATAGUISE DGSECURE: SAFEGUARDING SENSITIVE DATA IN THE CLOUD

Organizations recognize the benefits of employing an elastic and scalable public cloud infrastructure. But many, particularly those in highly regulated industries such as financial services and healthcare, feel compelled to keep sensitive data on-premises to meet privacy and regulatory mandates. Dataguise enables organizations to keep confidential information safe as they leverage the economics and scalability of the cloud, unlock value from big data, and derive benefits from enterprise-grade cloud services.

The Dataguise solution is security for the data that matters most: sensitive data. It precisely and continuously identifies, protects, and inspects data at a granular level, wherever it lives or moves across the enterprise and cloud, including big data repositories. DgSecure can discover sensitive data regardless of its data type or location, protect it via masking and/or encryption, define its relationship to other data, prevent unauthorized access, and provide continuous data monitoring with precise alerts. The result is improved risk management, flexible data sharing, and precise visibility into sensitive information in large pools of structured and unstructured data, such as clickstreams, logs, user documents, and email.

The Dataguise DgSecure solution enables organizations to determine where, how much, and how often sensitive data elements—credit card numbers, Social Security numbers, names, email addresses, health records, financial performance results, and more—appear across the entire Hadoop data store. By using DgSecure auditing functions, organizations can see who is accessing data, how frequently they are accessing it, and what are the results (e.g., read, write, delete, copy, failed attempts). DgSecure also secures data in-flight as it moves by data transfer tools such as FTP and Flume.

DgSecure offers end-to-end data security whether Hadoop clusters reside on-premises or in the cloud. Organizations can protect their Hadoop data repository in the Azure cloud by using the DgSecure application available in the Azure marketplace. In a hybrid deployment scenario, DgSecure can protect sensitive information on-premises before and during a data transfer to the cloud.

Sensitive Data Security in 5 Simple Steps

The Dataguise DgSecure platform makes it easy to detect, protect, and monitor sensitive data assets—continuously and automatically—wherever they live and move across the enterprise and the cloud.

1. Understand how data is collected, transferred, and stored across the extended enterprise.
2. Define policies for what data elements are considered sensitive or subject to regulatory requirements (e.g., HIPAA, PCI, PHI, PII).
3. Discover existing and continually detect new sensitive data across all repositories.
4. Automatically encrypt or mask sensitive data according to policies.
5. View, monitor, and alert on sensitive data activities and risks in real time.

HDINSIGHT AND DGSECURE: SAFELY UNLOCKING THE BENEFITS OF BIG DATA IN THE CLOUD

Azure HDInsight and DgSecure together enable organizations to leverage the many benefits of big data and the cloud without putting sensitive data at risk. Businesses can deploy high-availability Hadoop distributions in the cloud with complete end-to-end security where the environments can scale elastically on demand.

With DgSecure, Azure users can securely move data between on-premises and cloud infrastructure with full workflow and automation, simplifying security across data repositories. Once identified and analyzed, the data can be protected at the element level by using masking or encryption and then democratized for use by business innovators.

Azure HDInsight customers can use DgSecure to detect where sensitive data resides across Hadoop deployments of any size. The DgSecure auditing functions provide visibility on what sensitive data is connected to and co-mingled with, as well as who is accessing it. Additionally, DgSecure can continuously monitor Azure HDInsight repositories to prevent unauthorized access.

Together, HDInsight and DgSecure open the possibilities for organizations to maximize the business value of their data—all of their data—including the valuable and sensitive data that was previously inaccessible to business users, development teams, and partners.

Consider these opportunities:

- **Maximize competitive advantage:** Gain better business intelligence and uncover new business opportunities with availability of data that was previously inaccessible
- **Reduce risks:** Adhere to data privacy and compliance regulations; understand sensitive data exposure and minimize the legal, financial, and brand implications
- **Accelerate time to value:** Enable greater access to more accurate and complete data faster to speed innovation and improve quality
- **Reduce costs:** Eliminate manual sensitive data discovery, compliance violation fees, and additional infrastructure requirements

HDInsight and Dataguise

"With Azure HDInsight you can spin up a Hadoop cluster in minutes and deploy in Windows or Linux. Dataguise DgSecure scales to meet user requirements for security without impacting functionality or performance, allowing users to gain the full value of this solution."

—Lance Olson, Partner Group Program Manager for HDInsight

"With the integration of Dataguise DgSecure and Microsoft Azure, enterprises can move to the cloud with complete confidence that their sensitive data will be better protected. DgSecure's unique, data-centric approach to on-premises and cloud-based data security enables seamless deployments to Microsoft Azure with higher utilization rates and more productive user experiences."

—Nicole Herskowitz, Senior Director of Product Marketing, Microsoft Azure.

NEXT STEPS

Learn more about secure and compliant Hadoop deployments on Microsoft Azure

- Read more about the Dataguise DgSecure data-centric security platform.
- Explore the capabilities of DgSecure on Azure in a free trial.
- Get introduced to HDInsight on Azure.
- Get a free trial of Azure to test Hadoop in HDInsight.

DATAGUISE
SECURE BUSINESS EXECUTION

In partnership with:

Microsoft Azure