WHITE PAPER

The Big Data Security Gap: Protecting the Hadoop Cluster

Zettaset

Introduction

While the open source framework has enabled the footprint of Hadoop to logically expand, enterprise organizations face deployment and management challenges with big data. Hadoop's core specifications are still being developed by the Apache community and, thus far, do not adequately address enterprise requirements for robust security, policy enforcement, and regulatory compliance. While Hadoop may have its challenges, its approach, which allows for the distributed processing of large data sets across clusters of computers, represents the future of enterprise computing.

Hadoop and similar NoSQL data stores enable any organization, large or small, to collect, manage and analyze immense data sets, but these nascent technologies were not designed with comprehensive security in mind. The response from data security vendors, who provide solutions for traditional structured databases, has been to modify their existing off-the-shelf products to secure the cluster environment.

However well-intentioned these independent approaches may be, each lacks a complete and focused security solution for Hadoop. Only a new approach that addresses the unique architecture of distributed computing can meet the security requirements of the enterprise data center and the Hadoop cluster environment.

This paper reviews the security gaps that exist in all open source Hadoop distributions while exploring and evaluating the disparate paths to Hadoop security being taken by Hadoop distribution and data security vendors. Finally, a solid pathway for securing distributed computing environments in the enterprise is provided.

Big Data Presents a New Security Challenge

Big data originates from multiple sources including sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. Thanks to cloud computing and the socialization of the Internet, petabytes of unstructured data are created daily online and much of this information has an intrinsic business value if it can be captured and analyzed.

For example, mobile communications companies collect data from cell towers; oil and gas companies collect data from refinery sensors and seismic exploration; electric power utilities collect data from power plants and distribution systems. Businesses collect large amounts of user-generated data from prospects and customers including credit card numbers, social security numbers, data on buying habits and patterns of usage.

The influx of big data and the need to move this information throughout an organization has created a massive new target for hackers and other cybercriminals. This data, which was previously unusable by organizations is now highly valuable, is subject to privacy laws and compliance regulations, and must be protected.

Hadoop is Not a Secure Technology

Hadoop, like many open source technologies such as UNIX and TCP/IP, was not created with security in mind. Hadoop evolved from other open-source Apache projects, directed at building open source web search engines. Hadoop was a spin off sub-project of Apache Lucene and Nutch projects, which used a MapReduce facility and a distributed file system with no built-in security. Hadoop is also the open-source version of the Google MapReduce framework, and no security was designed into the software as the data being stored (public URLs) was not subject to privacy regulation.

The open source Hadoop community supports some security features through the current implementation of Kerberos, the use of firewalls, and basic HDFS permissions. Kerberos is not a mandatory requirement for a Hadoop cluster, making it possible to run entire clusters without deploying any security. Kerberos is also difficult to install and configure on the cluster, and to integrate with Active Directory (AD) and Lightweight Directory Access Protocol, (LDAP) services. This makes security problematic to deploy, and thus constrains the adoption of even the most basic security functions for users of Hadoop.

"Hadoop, like many open source technologies such as UNIX and TCP/IP, was not created with security in mind."

Enterprise organizations have been subjected to the risks associated with data security breaches for decades now, and expect that any new technology that is adopted by IT and installed in the datacenter will meet a minimum set of security requirements. Enterprises want the same security capabilities for big data as that in place for "non-big data" information systems, including solutions that address user authentication and access control, policy enforcement and management, and data masking and encryption. Many organizations require these big data safeguards in order to maintain regulatory compliance with HIPAA, HITECH, SOX, PCI/DSS, and other security and privacy mandates.

To date, the open source community has not addressed these security gaps, and remains focused on creating improved Hadoop technologies such as MapReduce 2.0. For enterprise organizations with data at risk, especially those companies that must adhere to regulatory compliance mandates, this should be cause for concern.

Hadoop's Architecture Presents Unique Security Issues

Big data is distinguished by its fundamentally different deployment model: highly distributed, redundant, and elastic data repositories enabled by the Hadoop File System. Rather than being a siloed, centralized data repository, such as a solitary Oracle Database, a Hadoop cluster may consist of anywhere from tens to thousands of nodes. This group of machines works in tandem to appear as a single entity, much like a mainframe, but with much lower capital expense and operating cost. But the characteristics of Hadoop's distributed computing architecture present a unique set of challenges for datacenter managers and security professionals.

- Distributed computing - Data is processed anywhere resources are available, enabling massively parallel computation. This creates complicated environments that are highly vulnerable to attack, as opposed to the centralized repositories that are monolithic and easier to secure.
- Fragmented data - Data within big data clusters is fluid, with multiple copies moving to and from different nodes to ensure redundancy and resiliency. Data can become sliced into fragments that are shared across multiple servers. This fragmentation adds more complexity to the security challenge.
- Access to data - Role-Based Access Control (RBAC) is central to most database security frameworks, but most big data environments only offer access control at the schema level, with no finer granularity to address users by role and related access.
- Node-to-node communication - Hadoop and the vast majority of distributions don't communicate securely; they use RPC over TCP/IP.
- Virtually no security - Big data stacks build in almost no security. Aside from service-level authorization and web proxy capabilities from YARN, no facilities are available to protect data stores, applications, or core Hadoop features. All big data installations are built on the web services model, with few or no facilities for countering common web threats.

Why Traditional Security Solutions Fall Short

Since big data represents an alluring market opportunity, incumbent security vendors are trying to back into this space by reverse engineering their existing point solutions to work in a distributed environment. While there is recognition that big data requires a different approach to security, targeted solutions do not currently exist.

"The massive volume, velocity, and variety of data are overwhelming to existing security solutions which were not designed and built with big data in mind."

The massive volume, velocity, and variety of data are overwhelming to existing security solutions which were not designed and built with big data in mind. Hadoop is not a single technology, but an entire eco-system of applications including Hive, HBase, Zookeeper, Oozie, and Job Tracker. Each of these applications requires hardening. To add security capabilities into a big data environment, functions need to scale with the data. Bolt-on security does not scale well, and simply cannot keep up.

Security vendors have adapted their existing offerings as well as they can, generally applying a control point (gateway/perimeter) where data and commands enter the cluster. However, they do not apply security where it is most effective: within the cluster.

Why Perimeter Security Leaves Data Center Clusters Exposed

Incumbent data security vendors believe that Hadoop and distributed cluster security can be addressed with traditional perimeter security solutions such as firewalls and intrusion detection/prevention technologies. But no matter how advanced, traditional approaches that rely on perimeter security are unable to adequately secure Hadoop clusters and distributed file systems.

Some firewalls attempt to map IP to actual AD credentials, but this is problematic in the Hadoop environment. In order to function accurately this requires specific network design (i.e., no NAT from internal corporate sub-nets). Even with special network configuration, a firewall can only restrict access on an IP/port basis, and knows nothing of the Hadoop File System or Hadoop itself.

As a result, data administrators have to segregate sensitive data on separate servers in order to control access. This approach is fundamentally incompatible with distributed file systems like Hadoop. It would require the creation of a second Hadoop cluster to contain sensitive data, and even then would only provide two levels of security for the data.

As depicted in Figure 1, even without those drawbacks, firewalls cannot adequately address Hadoop security. The problem with all firewalls is that they represent a single layer of defense around a soft interior. Once a firewall is breached, the cluster is wide open for attack. Firewalls offer no protection for data at-rest or in-transit within the cluster.

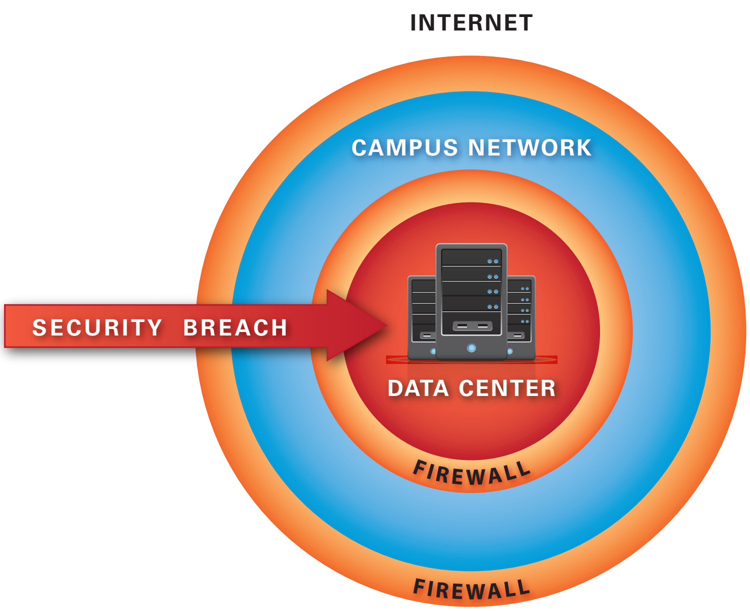

Figure 1: Firewalls and other perimeter security solutions represent a single layer of defense that, when breached, offer no real protection for the data cluster.

Firewalls also offer no protection from breaches which originate from within the firewall perimeter. An attacker who can get into the data center either physically or electronically can steal whatever they want, since the data is un-encrypted and there is no authentication required for access.

Solution: Move Security Closer to the Data

A Forrester report, the "Future of Data Security and Privacy: Controlling Big Data", observes that security professionals apply most controls at the very edges of the network. However, if attackers penetrate your perimeter, they will have full and unrestricted access to your big data. The report recommends placing controls as close as possible to the data store and the data itself, in order to create a more effective line of defense. Thus, if the priority is data security, then the cluster must be highly secured against attacks.

Deploy a Purpose-Built Security Solution for Hadoop and Big Data

Only a new approach that addresses the unique architecture of distributed computing can meet the security requirements of the enterprise data center and the Hadoop cluster environment.

"Only a new approach that addresses the unique architecture of distributed computing can meet the security requirements of the enterprise data center and the Hadoop cluster environment."

Zettaset Orchestrator provides an enterprise-class security solution for big data that is embedded in the data cluster itself, moving security as close to the data as possible, and providing protection that perimeter security devices such as firewalls cannot deliver.

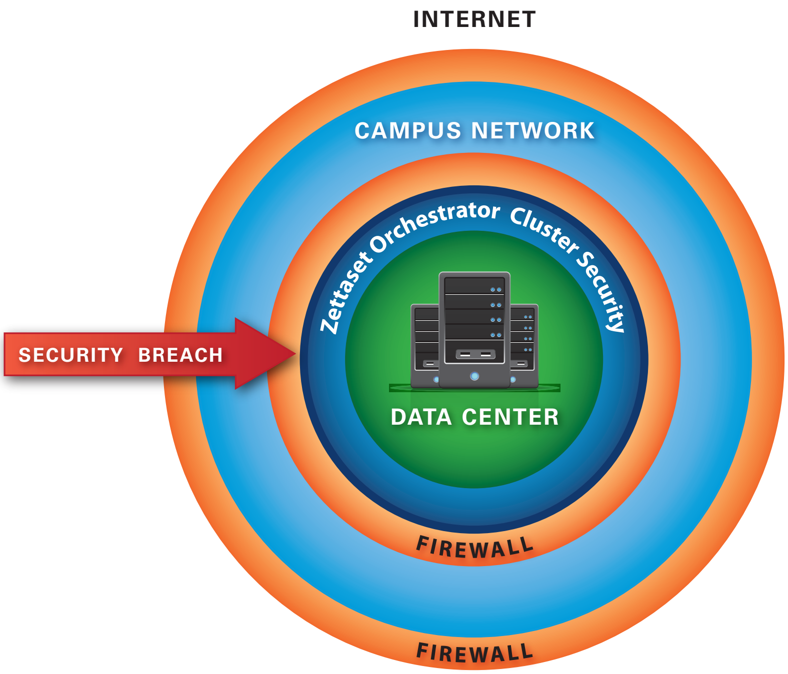

Figure 2: Zettaset Orchestrator provides security from within the data center cluster. Even if perimeter security is breached, the cluster and sensitive data are still protected by Orchestrator's comprehensive security wrapper.

At the same time, Orchestrator addresses the security gaps that open-source solutions typically ignore, with a comprehensive big data management solution which is hardened to address policy, compliance, access control and risk management within the Hadoop cluster environment.

Orchestrator includes RBAC, which significantly strengthens the user authentication process. Orchestrator simplifies the integration of Hadoop clusters into an existing security policy framework, with support for LDAP and AD. For those organizations with compliance reporting requirements, Orchestrator includes extensive logging, search, and auditing capabilities.

Orchestrator addresses the critical security gaps that exist in today's distributed big data environment with these capabilities:

- Fine-grained Access Control – Orchestrator significantly improves the user authentication process with RBAC.
- Policy Management – Orchestrator simplifies the integration of Hadoop clusters into an existing security policy framework with support for LDAP and AD.
- Compliance Support – Orchestrator enables Hadoop clusters to meet compliance requirements for reporting and forensics by providing centralized configuration management, logging, and auditing. This also enhances security by maintaining tight control of ingress and egress points in the cluster and history of access to data.

Zettaset Orchestrator is the only solution that has been specifically designed to meet the security requirements of the distributed architectures which predominate in big data and Hadoop environments. Orchestrator creates a security wrapper around any Hadoop distribution and distributed computing environment, making it enterprise-ready. With Orchestrator, organizations can now confidently deploy Hadoop in data center environments where security and compliance is a business imperative.

"Zettaset Orchestrator is the only solution that has been specifically designed to meet the security requirements of the distributed architectures which predominate in big data and Hadoop environments."

About Zettaset

Zettaset is an enterprise software company based in Mountain View, California, driven by technology industry leaders with experience in the enterprise software, security, and networking domains. Zettaset provides its critical enabling technology to end-user customers through a network of strategic partners.

Zettaset

465 Fairchild Drive, Suite 207, Mountain View, CA 94043 // www.zettaset.com
USA: +1.650.314.7920 // Fax: +1.650.314.7950 // sales@zettaset.com