DATA GOVERNANCE IMPLEMENTATION AND SECURITY THREAT ELIMINATION

Abstract

It is an unavoidable reality that large and damaging incidents of computer crime can be carried out by organization insiders. The assumption that “the bad guys” are on the outside is no longer the primary actuality concerning data security. It is imperative for all organizations to strengthen their data governance and security programs to protect sensitive and proprietary information while eliminating the potential for improper data access and utilization. The term “data governance” is a general term that covers the practices and policies organizations create and abide by to ensure proper management and utilization of their data. Data security is the process of protecting data from unauthorized access and corruption. To implement data governance and eliminate data security threats, the best-in-class tools for authentication, authorization, and integration are recommended: Kerberos, Apache Ranger, and Apache Atlas.

Problem Statement

IBM estimated that poor data quality cost the United States economy $3.1 trillion in 2016. The poor quality is due to a lack of data governance which leads to the creation of data security threats within an organization. It is crucial for organizations to understand the aspects of governance and security and then select the best tools to manage both issues.

Data Governance

Data governance is a confluence of policies and strategies which address the creation and usage of granular data as inputs into a system. Data governance is similar to information governance; the lifecycle management of the information derived from the data including its use, protection, and preservation. Therefore, due to the relationship between the data and information, information cannot be governed if data is not governed.

Currently, most organizations have little to no tracking of where data is coming from, who is using it, and where it is going. Shown in the graphic to the right are the results regarding designated data governance and management personnel within the surveyed organizations from the 2018 Digital Analytics and Data Governance Report as conducted by ObservePoint.

This report surveyed 546 industry professionals of various verticals and backgrounds in effort to examine how organizations govern and analyze data. As the figure illustrates how organizations are still struggling executing data governance with 17% of respondents unsure if their organizations have designated data governance personnel and 28% of respondents stating that no data governance personnel were in their organization.

Organizations must acknowledge their antiquated data governance processes. Most concerning are the organizations which have absolutely no data governance practices. This lethargic stance on data creates a massive legal and financial risk. With the democratization of data science, utilization and interpretation of data is no longer the sole responsibility of the data scientists. As the data sources become more open to more departments, the data governance policies and practices should continuously become more sophisticated. Without a defined set of practices, policies, standards, and guideposts, an organization will fail to meet current performance objectives and future successes.

Organizations need audits to determine who did what with which data. Auditability paired with data governance reduces the threat of legal risks and misinterpretations of data. Organizations who maintain proper data governance will benefit from operational efficiency, enhanced understanding of data, greater data quality, and enlightened decision-making, which will ultimately lead to increased revenue.

Data Security

Data Security is the process of protecting data, especially data containing personal information, from unauthorized access and corruption, both internally and externally. As defined by the Personal Information Protection and Electronic Documents Act (PIPEDA), “personal information includes any factual or subjective information, recorded or not, about an identifiable individual.” In order to mitigate harm to all data, organizations deploy data encryption, tokenization, and key management practices.

Data Security
5 Key Principles
1. Take Stock
2. Scale Down
3. Lock It
4. Pitch It
5. Plan Ahead

In accordance with the United States Federal Trade Commission (FTC), a sound data security plan applies five key principles.

- The first principle highlights data governance through the importance of inventorying personal information within the organization’s data structure as well as the identification of personnel with access to this information.
- Scaling down the amount of personal information kept within the data structure is the second principle. Therefore, if the personal information is not essential for business operations, it should not be retained, or even initially gathered.
- The third principle is to protect the information that is kept within the organization, both physically and electronically. Information protection should be conducted through many channels including, but not limited to, encryption, anti-malware software, restriction of downloads of unauthorized software, disabling unused ports, establishing authentication protocols, utilizing firewalls, and breach detection.

- Disposal of personal information which is no longer needed for business purposes is the fourth key principle of data security. This is not purely specific to information stored within the organization's data structure. Computers, phones, tablets, and the like, which are no longer used by the organization, should have their data erased through a wipe utility program before disposal.
- The fifth and final principle is to create a plan for responding to security incidents. Should a security incident arise, a senior member of the staff should act to implement a pre-determined response plan which can include network disconnection of the insecure device, investigation procedures, and proper notification protocol for all need-to-know parties.

Solutions for Data Governance and Security

Authentication: Kerberos

Overview

Authentication is the foundation for proper data governance and security. Developed by the Massachusetts Institute of Technology, the Kerberos protocol is freely available with copyright permissions similar to those used for the X Window System and Berkeley Software Distribution operating system. Kerberos is also available from vendors who provide professional implementation and support of the product.

Functionality

Kerberos is a network authentication protocol which utilizes a secret-key cryptography to provide strong authentication for client/server applications. Kerberos encryption occurs between a client's identity to a server, or from a server to a client, across an insecure network connection. This encryption can also provide a secure communications protocol to provide data integrity and privacy. The Kerberos "single sign-on" authentication and authorization is a universal solution for data security.

Authorization: Apache Ranger

Overview

Upon authentication with Kerberos, the user's access rights must be determined. Apache Ranger is the framework in which this process occurs. Simultaneously, Ranger is auditing and managing all of the comprehensive data security across the Apache Hadoop platform. Through Ranger, controls are established concerning a user's access rights to specific resources. Resources (file, folders, databases, tables, columns, rows, etc.) are easily managed through the generation of policies for a particular set of users and/or groups.

Apache Ranger also provides security through the delegation of data administration to specific group owners. This allows for the decentralization of data ownership throughout the organization. Ranger allows for the creation of services for specific Hadoop resources (HDFS, HBase, Hive, etc.) and the addition of access policies to those services. Generation of access policies through tag-based services can then be applied to the specific Hadoop services. Utilization of tag-based policies enables controlled access to resources across multiple Hadoop components without creating separate services and policies within each resource component. Finally, and most importantly, Apache Ranger provides the ability to enable audit tracking and policy analytics for a deeper control of the environment.

Apache Ranger Features			
Administration Console	**Centralized Web Application**	**NameNode**	**Security**
Enable tracking and policy analytics for control of the data environment	Enforce security policies within Hadoop ecosystem using lightweight Ranger Java plugins	Plugin through which the user makes a request and a decision is reached if the request should be authorized	Enforce authorization using native Hadoop file-level permissions if Ranger policies do not cover necessary access
Manage policies for specific users and groups	Allows for authorized users to manage security policies using web tools or APIs	Collects access request details required for auditing	Enforce previously generated security policies

Functionality

Apache Ranger contains a centralized web application for the management of policy administration, auditing, and reporting modules. This centralized security framework enables fine grained access control over Hadoop and its related components. If a policy server goes down temporarily, the Ranger plugin will continue to function and provide authorization enforcement.

To create and manage authorized users, the Representational State Transfer Application Programming Interface (RESTful API) is managed through the web application. Security policies are also managed through this application. This is where the security administrator can base policies on pre-defined data classifications such as personally identifiable information (PII), customer proprietary network information (CPNI), or information covered within the Health Insurance Portability and Accountability Act (HIPAA). The Tag Synchronization Module (TagSync) allows for coordination between the Ranger security administrator and the metadata tag-source, Apache Atlas.

Metadata Management: Apache Atlas

Overview

Apache Atlas is an exclusive tool for the Apache Hadoop platform. This asynchronous tool provides governance capabilities and metadata management for organizations. The exchange of metadata through Atlas, inside and outside of Hadoop, enables organizations to have a true platform-agonistic governance which adheres to the strictest compliance requirements. Metadata is managed with the application of types, entities, and attributes. An entity's definition of an attribute value must match a multiplicity property. If not, a constraint violation occurs, and the entity addition fails. This violation ensures the integrity of the data.

Apache Atlas is scalable in building, classifying, and governing a catalog of data assets. Additionally, up-to-date perspectives on the data are available through change notifications within the data landscape. New and changed data sources can trigger automated metadata discovery to assist in the generation of a rich definition of the metadata repository. Two main mechanisms serve this process: bridges and hooks. The bridge is the initial load of metadata from the data platform, service, or engine. The hook is the continuous feed of resource changes to Atlas.

Apache Atlas Features					
Knowledge Store	**Data Lifecycle Management**	**Audit Store**	**Security**	**Policy Engine**	**RESTful Interface**
Categorize data into a taxonomy of sets, objects, tables, and columns	Leverage existing feed processing and management system	Historical repository for all governance, security, and operational events	Establish global security policies based on data classifications	Fully extensible policy engine to allow for the addition of new capabilities	Supports extensibility through REST APIs
Support the exchange of metadata between HDP and third-party applications	Focused on provenance, multi-cluster replication, data set retention and eviction, late data handling, and automation	Indexed and searchable	Integrates with HDP through Apache Ranger plug-in for security policy enforcement	Supports rationalization runtime rules: metadata-, geo-, and time-based	Allows for third-party applications to use existing tools to view and update metadata

Functionality

To represent the managed metadata objects, Atlas uses a JanusGraph model. JanusGraph is a transactional database which is scalable for optimal storing and querying of graphs which can contain hundreds of billions of edges and vertices. Distributed across multi-machine clusters, JanusGraph can support thousands of concurrent users rendering intricate graph traversals in real time.

Specifically for Atlas, the JanusGraph repository shows the interconnected relationships between data sources, the hosted data sets, the business meaning of data elements, and the classification of these elements. Classification is based upon quality, confidentiality, and retention. The Atlas model builds upon the interconnected metadata relationships by building out specific structures for their storage. The metadata types are defined through calls to the Types API or through JSON files.

Atlas utilizes a model for users to define the metadata object they wish to manage. The model contains types which are definitions of how particular kinds of metadata objects are stored and accessed. A type is similar to a class or table schema in that a type represents one or a collection of attributes that define the metadata object. The following categories are types within the system: entity, classification, relationship, struct, and enumeration.

Specific instances of types are called entities. Entities represent the actual managed metadata objects. These entities can be associated with multiple classifications and can include dynamic security designations. Based on the entity-type, -classification, or -id, the authorization model enables control of which users and/or groups can perform the following operations: read, create, update, delete, read classification, add classification, update classification, and remove classification. Additional administrator operations are available to allow users and/or groups to import entities and export entities without entity level access.

The final component of the model is the attribute. The attribute is used to influence the specific modeling behavior required by Atlas. Attributes must have the following properties: attribute name, metatype name, isComposite, isIndexable, isUnique, and multiplicity. Regarding metadata constraints, the multiplicity attribute is used to indicate if the attribute is required, optional, or multi-valued. A violation will occur if the multiplicity declaration is not matched between an entity definition and attribute value. As previously stated, this violation ensures the integrity of the data.

Integration: Atlas-Ranger

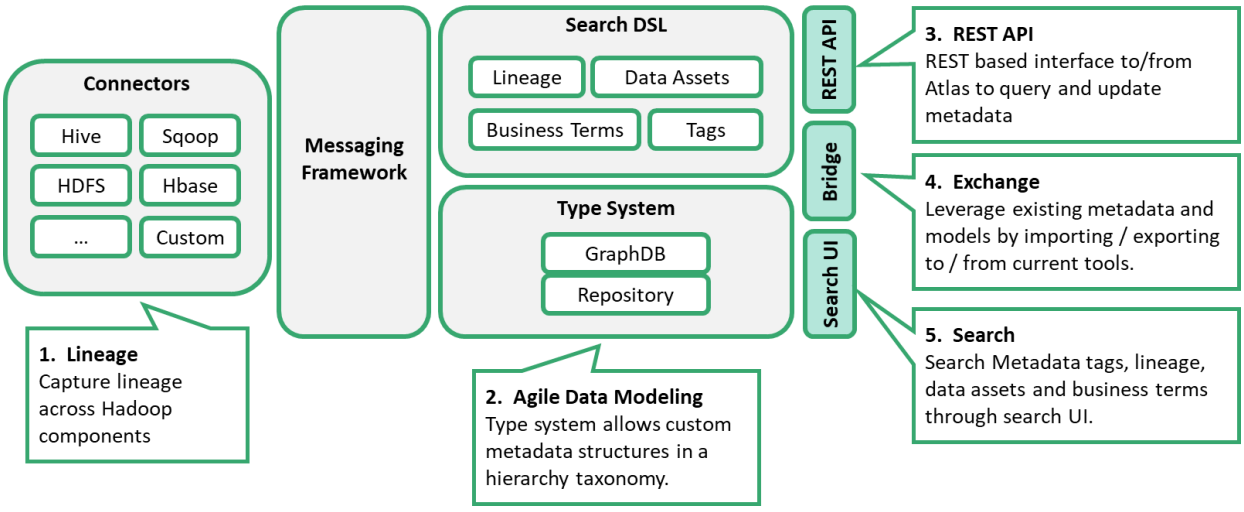
Overview

When integrated with Apache Ranger, Atlas will enable authorization and data-masking based on three resource hierarchies: types (create/update/delete any classification type), entity (perform all operations on metadata entities), and admin (export/import). This integration is key to unite the data classification and metadata store capabilities of Atlas with the security enforcement of Ranger. To define and dynamically implement security policies, the attribute-based tags created within the application are also incorporated within Ranger.

Functionality

Four access policies are generated with the merger of Ranger and Atlas. The first is classification-based access controls in which an entity is marked with a metadata tag related to compliance or business taxonomy. The second is a data expiry-based policy; here data is tagged with an expiration date for business usage. When the expiration date has been reached, Ranger automatically denies access to the tagged data. The third policy grants or denies access based on location-specific constraints. For this policy, privacy rules are evaluated based on the user's geographic location at the time of data request. The fourth policy is prohibition against dataset combinations which prevents the violation of combined data queries due to security policies.

The following graphic displays the incorporation of Apache Atlas and Ranger within the Hadoop platform. Five key functions regarding the data governance and security capabilities within the platform are pointed out within the graphic.



Solution Implementation

Due to the lack of data governance within many organizations, data is liable to internal security threats and legal as well as financial risks. Proper data governance requires the utilization of consistent metadata architecture through tagging of data types to protect sensitive information. Paired with administration of user and group access policies and the generation of data lineage records, data security is ensured within an organization. To implement data governance and eliminate data security threats, the best-in-class tools for authentication, authorization, and integration are recommended: Kerberos, Apache Ranger, and Apache Atlas. Oalva, Inc. is available to guide, assist, and implement these tools to ensure the proper process of data governance and the installation of stringent data security protocols.

For more information on how to implement this solution, please contact Oalva, Inc. at info@oalva.com.

About the Authors

Margaret Baer

Margaret Baer is a Big Data and A.I. Solutions Specialist at Oalva, a leading Hadoop solutions integration company in the United States and Canada. Margaret will graduate from Utica College in May 2019 as a Master of Science in data science. Margaret is passionate about improving the lives of others through the implementation of data science.

Tim Fox

Tim Fox is a Big Data and Hadoop Specialist at Oalva. Tim is currently attaining his Bachelor of Science in computer science degree from the University of Kansas. Tim is driven to find the best solutions for organizations' needs through big data management.

Special Thanks To:

James Dinkel

President, Oalva, Inc.

Michael McCarthy, Ph.D.

Assistant Professor and Director of Data Science, Utica College

Roxie Baer

Chief Editor

References

American Health Information Management Association. (2016, February 3). Information Governance FAQs. Retrieved from http://www.ahima.org/topics/infogovernance/faqs?tabid=faq

Hortonworks, Inc. (2018, April 1). Data Governance Overview. Retrieved from https://docs.hortonworks.com/HDPDocuments/HDP3/HDP-3.0.0/atlas-overview/content/apache_atlas_features.html

Massachusetts Institute of Technology. (2019, January 9). Kerberos: The Network Authentication Protocol. Retrieved from https://web.mit.edu/kerberos/

Bean, R. (2016, November 08). The Case For 'Data Governance'. Retrieved from https://www.forbes.com/sites/ciocentral/2016/06/22/the-case-for-data-governance/#4bdf965a54be

Chessell, M. (2017, July 14). Apache Software Foundation. Retrieved from https://cwiki.apache.org/confluence/display/ATLAS/Atlas Model

Chessell, M. (2017, May 16). Apache Software Foundation. Retrieved from https://cwiki.apache.org/confluence/display/ATLAS/Atlas Bridges and Hooks

Chessell, M. (2018, March 30). Apache Software Foundation. Retrieved from https://cwiki.apache.org/confluence/display/ATLAS/Atlas Graph Strategy

Ching, M. O. (2019, January 3). Apache Ranger – Introduction. Retrieved from https://ranger.apache.org/

Cole, Z. (2017, December 7). The Top 6 Benefits of Data Governance. Retrieved from https://erwin.com/blog/top-6-benefits-of-data-governance/

Cornelissen, J. (2018, September 18). The Democratization of Data Science. Retrieved from https://hbr.org/2018/07/the-democratization-of-data-science

Federal Trade Commission. (2017, September 06). Protecting Personal Information: A Guide for Business. Retrieved from https://www.ftc.gov/tips-advice/business-center/guidance/protecting-personal-information-guide-business

Gupta, K. (2017, February 20). What You Need to Know About Apache Ranger. Retrieved from https://www.freelancinggig.com/blog/2017/02/20/need-know-apache-ranger/

Hallenbeck, C. (2018, June 1). The Vexing Problem Of Data Entropy. Retrieved from https://www.digitalistmag.com/cio-knowledge/2018/06/01/vexing-problem-of-data-entropy-06173208

Hallenbeck, C. (2018, June 14). The Challenge Of Data Sprawl: Accessibility And Quality. Retrieved from https://www.digitalistmag.com/cio-knowledge/2018/06/14/challenge-of-data-sprawl-accessibility-quality-06176426

Hallenbeck, C. (2018, June 21). Why Data Security And Data Governance Matter. Retrieved from https://www.digitalistmag.com/cio-knowledge/2018/06/21/why-data-security-data-governance-matter-06177299

IBM. (2016, January 11). Extracting business value from the 4 V's of big data. Retrieved from https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data

INFINITIVE DIFFERENCE BLOG. (2017, May 5). 5 Reasons Why Data Governance Matters to Data-Driven Marketers. Retrieved from https://www.infinitive.com/2017/05/05/5-reasons-why-data-governance-matters-to-data-driven-marketers/

JanusGraph Authors. (2017, June 13). JanusGraph. Retrieved from http://janusgraph.org/

Kulkarni, A. (2016, April 21). Apache Software Foundation. Retrieved from https://cwiki.apache.org/confluence/display/RANGER/Tag Synchronizer Installation and Configuration

Matthews, K. (2018, March 28). 10 data statistics information managers need to know. Retrieved from https://www.information-management.com/slideshow/10-data-statistics-information-managers-need-to-know

Micro Focus. (2018, August 18). What is Data Security? Retrieved from https://www.microfocus.com/en-us/what-is/data-security

MIT Kerberos & Internet Trust (KIT) Consortium. (2015). MIT Consortium for Kerberos and Internet Trust. Retrieved from http://kit.mit.edu/about/history

MuleSoft. (2017, November 20). What is REST API Design? Retrieved from https://www.mulesoft.com/resources/api/what-is-rest-api-design

ObservePoint. (2018, December 20). 2018 Digital Analytics & Data Governance Report: The Top 7 Insights. Retrieved from https://resources.observepoint.com/blog/2018-digital-analytics-data-governance-survey-top-7-insights

Office of the Privacy Commissioner of Canada. (2018, January 09). PIPEDA in brief. Retrieved from https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/

PwC. (2016, March 15). *Data Governance Survey Results: A European Comparison of Data Management Capabilities in Banks*(Rep.). Retrieved from https://www.pwc.fr/fr/assets/files/pdf/2016/05/pwc_a4_data_governance_results.pdf

Rockendorf, B. (2017, October 27). Data Governance is Imperative for Big Data Analytics. Retrieved from http://journal.ahima.org/2017/10/27/data-governance-is-imperative-for-big-data-analytics/

Rouse, M. (2005, April). What is CPNI (Customer Proprietary Network Information)? - Definition from WhatIs.com. Retrieved from https://searchnetworking.techtarget.com/definition/CPNI

Rouse, M. (2014, January). What is personally identifiable information (PII)? - Definition from WhatIs.com. Retrieved from https://searchfinancialsecurity.techtarget.com/definition/personally-identifiable-information

Rouse, M. (2019, February). What is HIPAA (Health Insurance Portability and Accountability Act) ? - Definition from WhatIs.com. Retrieved from https://searchhealthit.techtarget.com/definition/HIPAA

Sweeney, M., & Lubowicka, K. (2019, January 24). Personally Identifiable Information: What is PII, non-PII & Personal Data? Retrieved from https://piwik.pro/blog/what-is-pii-personal-data/

The Apache Software Foundation. (2016, July 9). Data Governance and Metadata framework for Hadoop. Retrieved from https://atlas.apache.org/0.7.0-incubating/

The Apache Software Foundation. (2018, September 18). Apache Atlas – Data Governance and Metadata framework for Hadoop. Retrieved from https://atlas.apache.org/

The Apache Software Foundation. (2018, September 18). Atlas Authorization Model. Retrieved from https://atlas.apache.org/Atlas-Authorization-Model.html

The Apache Software Foundation. (2018, September 18). Type System. Retrieved from https://atlas.apache.org/TypeSystem.html

The Apache Software Foundation. (2019, January 3). Apache Ranger – Frequently Asked Questions. Retrieved from https://ranger.apache.org/faq.html#What_does_Apache_Ranger_offer_for_Hadoop