IBM BigInsights Security Implementation: Part 1 Introduction to Security Architecture

Big data analytics involves processing large amounts of data that cannot be handled by conventional systems. The IBM BigInsights® platform processes large amounts of data by breaking the computation into smaller tasks that can be distributed onto several nodes. As this platform is shared by users in different roles (developers, analysts, data scientists, and testers), it introduces the challenge of provisioning access and authorization to the cluster and securing the data.

Big data platforms are an amalgamation of several individual components that are still evolving, and are based on the challenges and requirements that are dictated by the open source community. These components are developed in isolation by independent teams with no forethought of integrating them in a secure way, which results in individual components defining and exposing their own security policies for data and access protection. This inherent lack of single security policy enforcement in big data platforms can be challenging and overwhelming.

This IBM Redbooks® Analytics Support web doc introduces a reference security architecture for the IBM BigInsights solution that is in line with current industry practices. It can be used as a reference document for solution architects and solution implementers. This document applies to IBM BigInsights Version 4.2 and later.

Security aspects to consider when designing the security architecture for IBM BigInsights

Securing an IBM BigInsights cluster involves addressing four main security aspects, which are shown in Figure 1:

- Secure Perimeter
- Secure Data
- Access Management
- Audit

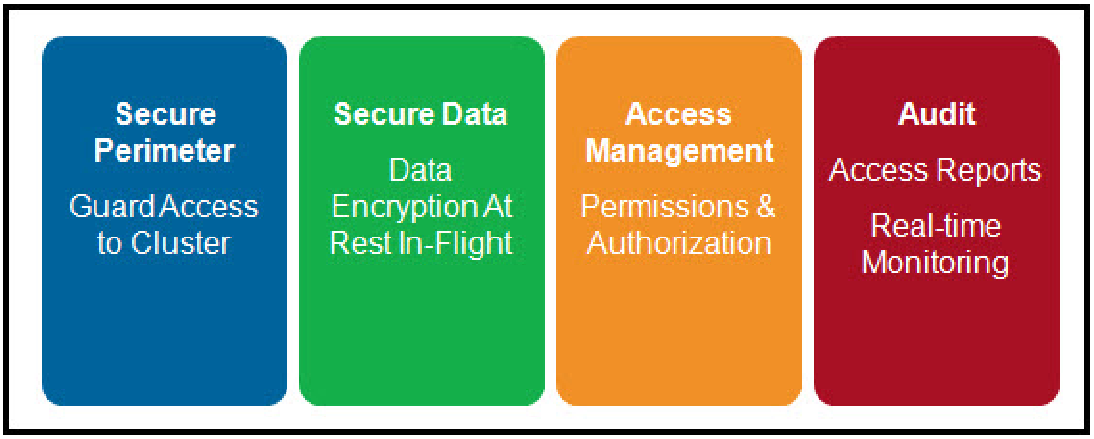

Figure 1. Four aspects of IBM BigInsights security

A secure perimeter can be enforced in the following ways:

- By authenticating users against LDAP and Kerberos
- By protecting HTTPS access through the Apache Knox security gateway
- By isolating the data nodes in a secure private network

Secure data can be accomplished in the following ways:

- By using Hadoop transparent encryption with Apache KMS (Key Management Server)
- By using IBM BigSQL data masking
- By enabling SSL and TLS support for components to secure the data transfers

Access management should be enforced at several levels:

- At the job level by using Yet Another Resource Scheduler (YARN) job-queue-based access control.
- By using SQL access privileges for SQL access of Hadoop data.
- By using ACL- based access control for Hadoop Distributed File System (HDFS) files.

Audits and reporting are provided by the following items:

- By using light-weight monitoring that uses Java Management Extensions (JMX)
- By using IBM Security Guardium® Data Activity Monitor

Figure 2 shows a high-level design of a secure IBM BigInsights cluster. It highlights various components that are the building blocks of a big data cluster architecture design.

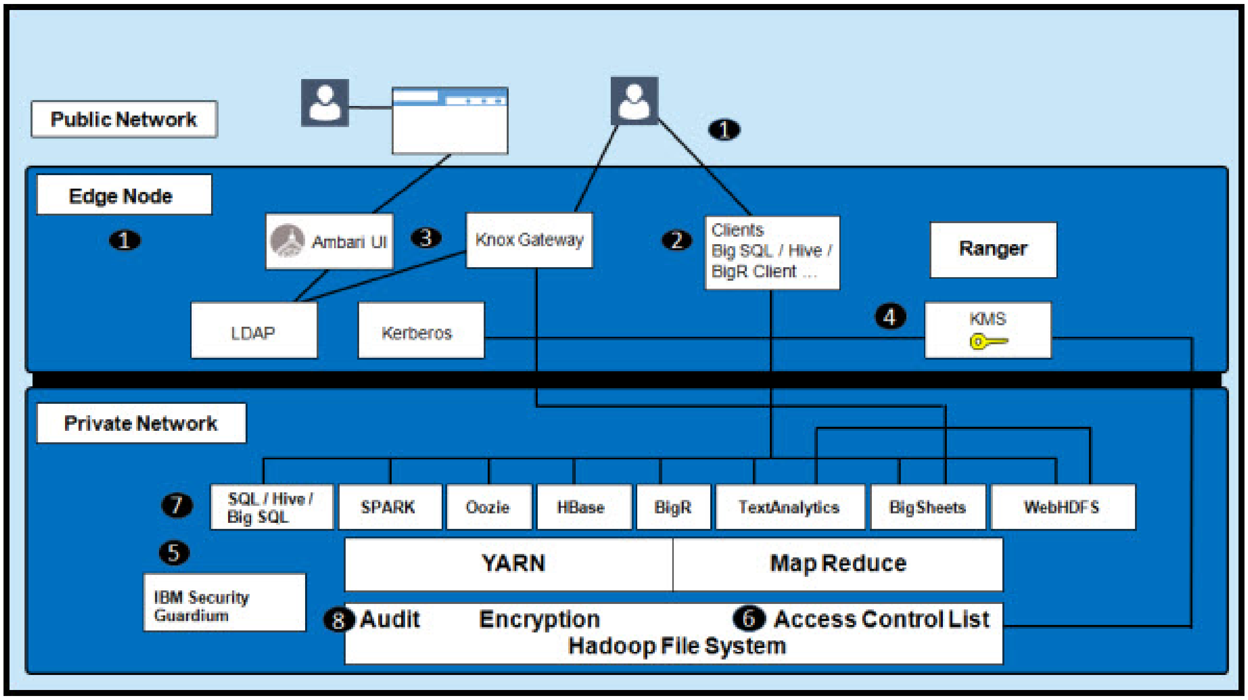

Figure 2. IBM BigInsights high-level design

Note the following items in Figure 2:

- An IBM BigInsights cluster can span over two networks: Public and private networks. The communication between the two networks occurs through an edge node (1).
- This edge node has the client components (2) for all the master services that are deployed in the cluster so that users can connect and perform analytics and administration. Access to administration and analytic tools is enforced through Ambari user management and the Knox gateway (3).
- Data encryption protects user data from unauthorized access and enforces industry security standards. Data can be encrypted at rest and while it is being transferred over the network. Encryption at rest is performed in two ways: By using Hadoop transparent data encryption and by using IBM Security Guardium Data Encryption.

  - Hadoop transparent data encryption uses the key management server (KMS), which holds encryption and decryption keys. (4)
  - IBM Security Guardium Data Encryption deploys agents on all nodes to perform encryption and decryption of data. The IBM Security Guardium server monitors and enforces encryption and decryption policies and rules on the agents. (5)

  The data transfers over the network are secured by configuring services to use SSL and TLS certificates.

- Similar to the Linux file system, Hadoop Distributed File System (HDFS) also provides fine-grained user access control by using file system access control lists (ACLs) (6). Big SQL and Hive provide Grant and Revoke commands to authorize users to perform certain operations (7).
- IBM Security Guardium Data Activity Monitor provides monitoring and auditing capabilities that you can use to integrate seamlessly Hadoop data protection into your existing enterprise data security strategy. HDFS has its own auditing mechanism that captures all the file system activities (8).

Designing the security architecture for IBM BigInsights products involves the features that are provided by individual components and a holistic approach that involves securing data, users, and functions from possible vulnerabilities.

Acknowledgements

Thanks to Mohan Dani, IBM BigInsights software developer, for his contributions to this project.

Related publications

- IBM BigInsights V4.2 documentation:
  https://ibm.biz/BdrnVB

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

This document was created or updated on August 25, 2016.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1340.html .

Trademarks