IBM BigInsights Security Implementation: Part 2 Securing the IBM BigInsights Cluster Perimeter

Big data analytics is network intensive because it runs on a cluster of nodes. Due to the high volumes of data exchanged between nodes in the IBM® BigInsights® cluster, network isolation is vital for the following reasons:

- To prevent sniffing attacks
- To reduce the network congestion so that the corporate network is not affected by the big data cluster traffic

Isolating the cluster network also gives administrators greater flexibility in enforcing the cluster access.

This IBM® Redbooks® Analytics Support web doc serves as a guide for system implementers who are creating a secure zone for an IBM BigInsights cluster by providing an example of current industry practices. This document applies to IBM BigInsights Version 4.2 and later.

IBM BigInsights cluster network deployment architecture

Figure 1 highlights one of the industry practices for network topology for IBM® BigInsights® clusters.

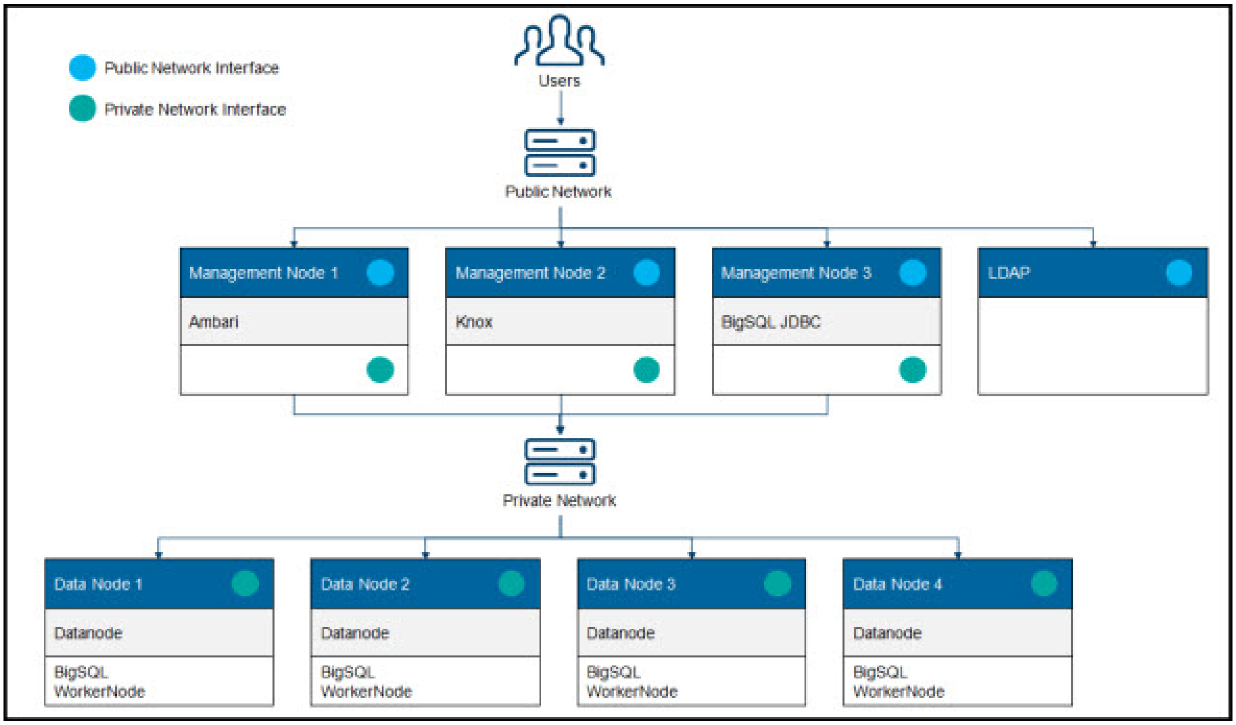

Figure 1. High-level IBM BigInsights network architecture

Users can access management nodes from the corporate network only after they authenticate with the corporate LDAP. Inbound traffic is encrypted and controlled by a firewall. Only ports that are related to cluster administration (Ambari), reverse proxy (Knox), JDBC ports (BigSQL and Hive), and SSH are open for users. Outbound traffic is not restricted. After users log in to the management node, they can connect via Secure Shell (SSH) to data nodes that are connected to the private network. All inbound traffic from the corporate network to data nodes is blocked. Data nodes are connected only to the cluster data network or to the private network.

Management nodes in the cluster have two network interfaces:

- One of the interfaces is connected to the corporate network (also called a *public network*).
- The other interface is connected to the private network (sometimes referred to as a *data network*).

All the data traffic is exchanged between management node and data nodes through the private network only. Thus, there is dedicated bandwidth and higher performance with the added benefit of security.

Setting up the data nodes to access external data sources with Apache Sqoop and similar tools

One of the use cases for big data analytics is to offload the organization's relational database management system (RDBMS) data to Apache Hadoop for archival or running analytics at large scale. Tools such as Apache Sqoop and IBM Fluid Query are used to import data from external sources. These tools launch MapReduce jobs, which read the data from external sources in parallel.

In this scenario, port forwarding in the firewall must be enabled, so that the data nodes can read external sources by forwarding the traffic to or from management nodes.

Run the following commands as root on the management nodes to enable port forwarding between data nodes and management nodes. Data nodes can then initiate communication to servers outside of the private network and receive data, but external servers cannot access the internal network.

```
echo 1 > /proc/sys/net/ipv4/ip_foward
/sbin/iptables -t nat -A POSTROUTING -o eth0 -j MASQUERADE
/sbin/iptables -A FORWARD -i eth0 -o eth1 -m state --state RELATED,ESTABLISHED 
-j ACCEPT
/sbin/iptables -A FORWARD -i eth1 -o eth0 -j ACCEPT
```

These commands assume `eth0` is the public interface. The `eth1` traffic is routed to outside of the network as though it were coming from the management node. Adjust the interface names (`eth0` and `eth1`) to match your environment. After the data is completely imported, port forwarding can be turned off.

Related publications

For more information, see the following web page:

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 December 8, 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/tips1348.html .

Trademarks