Kerberos: A Network Authentication for Hadoop

Shivani Beri[1], Priyanka Chaudhari[2], Rohit Kadam[3], Pranav Nimbalkar[4]

C S Mankarl[5]

[1] Student, Dept of Computer, MMIT, Maharashtra, India
[2] Student, Dept of Computer, MMIT, Maharashtra, India
[3] Student, Dept of Computer, MMIT, Maharashtra, India
[4] Student, Dept of Computer, MMIT, Maharashtra, India
[5] Professor, Dept of Computer, MMIT, Maharashtra, India

ABSTRACT

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data production is growing rapidly every year and as a adoption of this, Apache Hadoop brings out the authentication and authorization issue of the data resource security. We discuss an innovation solution to implement the token authentication based on the Kerberos pre-authentication framework. We put forth a pre authentication mechanism for Kerberos that allows to have a centralized access control for the platform. Users authenticate to Key Distribution Center (KDC) using a standardized token, and develop a plug-in for MIT Kerberos that can be implemented separately to employ the new mechanism. Depending on this, we propose a solution of token authentication for whole Hadoop data base which will make user authentication easier and more secure.

Cloud service providers are offering the popular Hadoop analytics platform following an "–as-a-service" model, i.e. clusters of machines in their cloud infrastructures pre-configured with Hadoop software. Such offerings lower the cost and complexity of deploying a comparable system on-premises, however security considerations and in particular data confidentiality hamper wider adoption of such services by enterprises that handle data of sensitive nature. In this paper, we describe our efforts in providing security for data-at-rest (i.e. data that is stored) when Hadoop is offered as a cloud service. We analyze the requirements and architecture for such service and further describe a new distributed file system that we developed for Hadoop called SDFS, towards supporting this premise. We analyze parameter tuning for SDFS and through experiments on a real test-bed we evaluate its performance. We further present simulation results that explore the parameter space and can guide tuning.

Keyword : - *Hadoop , Kerberos, KDC [Key Distribution Centre].*

1. INTRODUCTION

In an open network computing environment, a work station cannot be trusted to identify its users correctly to network services. Kerberos provides an alternative approach where by a trusted third- party authentication service is used to verify users 'identities. This paper gives an overview of the Kerberos authentication model as implemented for MIT's Project Athena. The Kerberos protocol can provide the following security functionality .Secure authentication based upon smartcards and public key certificates, RSAs Secure ID token and other authentication mechanisms. Having a single sign-on to kerberised applications and Cryptographic strength data integrity**.**

Cloud computing infrastructures have emerged as ideal platforms for running Big Data applications. They offer unprecedented scalability, flexible, pay-per-use charging models, and automatic provisioning of (physical or virtual) servers that enable end users to order and provision on- demand, high performance clusters pre-configured with a variety of Big Data platform software at a fraction of the cost and time that would be needed to deploy a comparable, on-premises solution. Figure 1 depicts a typical use case for Hadoop offered as a cloud service, which the main focus of this paper: enterprise users store their data in an enterprise storage server located in their premises. This data is regularly backed up and accumulated over time in public cloud storage. On a periodic basis, analysis on this data needs to take place, for example for quarterly financial forecasting purposes, and a temporary Hadoop cluster is provisioned on-demand in the cloud service provider. The enterprise data is moved from their long-term cloud storage to that cluster and analyzed. Once the analysis is completed, the results are moved back to the cloud storage service and are then synchronized with the enterprise, on-premises storage server, while the Hadoop cluster in the cloud is de- commissioned to save on costs till the next time it will be needed.

1.1 Security Concern

Security concerns hamper widespread adoption of the “-as-a-Service” model to Big Data analytics, especially for enterprises that handle sensitive data such as

financial and healthcare records. From the point of view of the enterprise, the on-premises storage server is considered to be a trusted environment, but neither the public cloud storage service nor the temporarily- provisioned Hadoop cluster in the cloud service provider is.

1.2 Encryption

Furthermore, while encryption may be employed to secure the data that resides in the cloud, it is of little use if the keys are stored and managed by the (public) cloud service provider, as the latter can still access the sensitive data stored by its customers. What is needed is an end-to-end approach for securing the enterprise data throughout the lifecycle of the analytics-as-a-service process that is controlled by the enterprise itself, as opposed to the cloud service provider.

2. LITERATURE SURVEY

[1] SDFS: Secure Distributed File System for Data-at-Rest Security for Hadoop-as a-Service.
AUTHOR :Petros Zerfos, Hangu Yeo, Brent D. Paulovicks and Vadim Sheinin IBM T.J.Watson Research Center York- town Heights, N Y U.S.A. pzerfos,hangu,ovicks,vadims@us.ibm.com .

The main efforts in providing security for data-at-rest (i.e. data that is stored) when Hadoop is offered as a cloud service. Analyze the requirements and architecture for such service and further describe a new distributed file-system that developed for Hadoop called SDFS, towards supporting this premise. Analyze parameter tuning for SDFS and through experiments on a real test-bed so evaluate its performance. Further present simulation results that explore the parameter space and can guide tuning. Data-at-rest security; Hadoop-as- a-service; secure distributed file system; Shamir’s secret sharing; information dispersal.

[2] A Token Authentication Solution for Hadoop Based on Kerberos Pre-Authentication

Author: Kai Zhen, Weihua Jiang Big Data Technologies Intel Corporation Shanghai , China

Kerberos protocol supports a pre-authentication framework that allows user to authenticate to KDC using other credentials instead of password via extended mechanisms. Developed a token solution based on Kerberos that avoids deployment overhead and risk as found in other solutions. Discuss how the Token Pre Authentication mechanism works for Kerberos, and then explained the application of it for Hadoop. Giving the advantages comparing with other similar solutions, we believe it provides a desirable option to integrate dominant identity and authorization providers like O Auth 2.0 for the ecosystem.

[3] Hadoop with Kerberos Architecture Considerations

Author: Stuart Rogers, Tom Keefer.

The non-secure configuration relies on client-side libraries to send the client-side credentials as determined from the client- side operating system as part of the protocol. While not secure, this configuration is sufficient for many deployments that rely on physical security. Authorization checks through ACLs and file permissions are still performed against the client- supplied user ID. After Kerberos is configured, Kerberos authentication is used to validate the client-side credentials. This means that the client must request a Service Ticket valid for the Hadoop environment and submit this Service Ticket as part of the client connection. Kerberos provides strong authentication in which tickets are exchanged between client and server. Validation is provided by a trusted third party in the form of the Kerberos Key Distribution Centre. To create a new Kerberos Key Distribution Centre specifically for the Hadoop environment, follow the standard instructions from the Cloudera or Horton works results.

[4]Distributed Authentication in Kerberos Using Public Key Cryptography.

Author: Marvin A.Sirbusirbu@cmu.eduJohnChung-IChuang chu ang + @ emu. edu Carnegie Mellon University Pitts- burgh,Pennsylvania15213

Fully distributed authentication using public key cryptography within the Kerberos ticket framework. By distributing most of the authentication workload away from the trusted intermediary and to the communicating parties, sign if cant enhancements to security and scalability can be achieved as comparedtoKerberosV5. Privacy of Kerberos clients is also enhanced. A working implementation of this extended protocol has been developed, and a migration plan is proposed for a transition from traditional to public key based Kerberos.

[5] Integrating Kerberos into Apache Hadoop

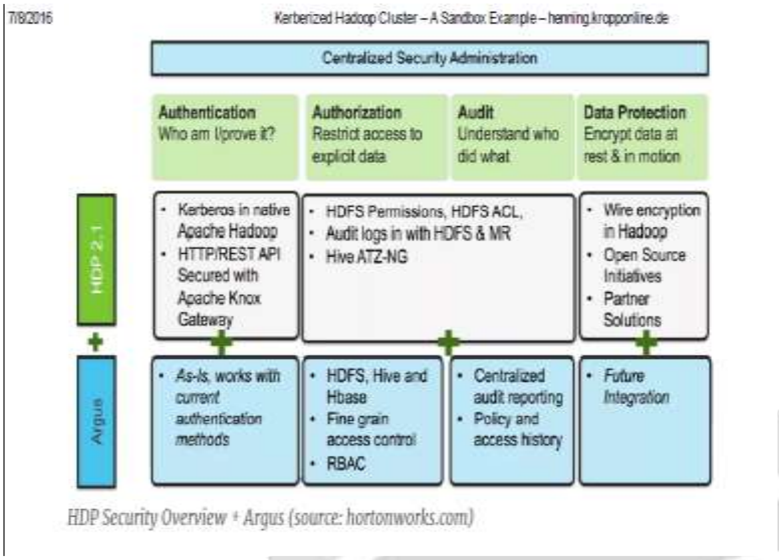
Author: Owen O Malley

HDFS: Communication between the client and the HDFS service is composed of two halves:

a) RPC connection from the client to the Name Node
b) Block transfer from the client to Data Nodes
c) The RPC connection can be authenticated via Kerberos or via a delegation token.
d) Map Reduce stores the information about the pending and executing jobs in HDFS, and therefore depends on HDFS remaining secure.

Accepts work flows over a HTTP inter- face that uses pluggable authentication. Oozie will run with the Kerberos service principal oozie and use that principal to authenticate to the HDFS and Map Reduce services. Both the HDFS and Map Reduce services will provide new methods that allow a super-user to act on behalf of others.

In an open network computing environment, a work station cannot be trusted to identify its users correctly to network services. Kerberos provides an alternative approach where by a trusted third party authentication service is used to verify users' identities. This paper gives an overview of the Kerberos authentication model as implemented for MIT' s Project Athena.



7/8/2016 Kerberized Hadoop Cluster – A Sandbox Example – henning.kropponline.de

Centralized Security Administration
Authentication
Who am I/prove it?
Authorization
Restrict access to explicit data
Audit
Understand who did what
Data Protection
Encrypt data at rest & in motion
HDP 2.1
• Kerberos in native Apache Hadoop
• HTTP/REST API Secured with Apache Knox Gateway
• HDFS Permissions, HDFS ACL,
• Audit logs in with HDFS & MR
• Hive ATZ-NG
• Wire encryption in Hadoop
• Open Source Initiatives
• Partner Solutions
Argus
• As-Is, works with current authentication methods
• HDFS, Hive and Hbase
• Fine grain access control
• RBAC
• Centralized audit reporting
• Policy and access history
• Future Integration

HDP Security Overview + Argus (source: hortonworks.com)

Fig -1: **System Architecture**

2.1 MATHEMATICAL MODEL

Let, 'S' be the system

- S={I, O, F, Su, Fa k Φs}

Where,

- I = set of input attributes.
- I ={B, login id, password, social media data(Implicit) k Φi }
- Where, B = set of requested information as input. B ={UUID, major id, minor id, proximity k Φ b}
- Φ b condition for beacon detection.
- Φ b: proximity >= 30m.
- Φ i: input condition
- Φ v:login id = valid,
- password = valid
- O = Set of Output attributes
- O ={Bi, Si k Φ o}

Where,

- Bi = Set of Book related in information
- Bi ={Summery, Cost, availability, recommendation k Φ Bi}
- ΦBi :Condition for displaying book in information.
- ΦBi :Book = available
- Si = Set of Statistical information for analysis
- Si ={Customer age grp, customer frq, top selling k ΦSi}
- Φ Si: Condition for statistical information .
- ΦSi :User information stored in database
- F = Set of necessary functions and additional functions.
- F ={Fr, Fp b, Fst}
- Where,
- Fr = functions that is used for finding recommendations.}
- Fpb= function for processing Beacon information if R< then intruders URL.

2.2 Algorithm to be used:

An Information Dispersal Algorithm (IDA) is developed that breaks a file, dispersal and reconstruction are computationally efficient. IDA has numerous applications to secure and reliable storage of information in computer networks and even on single disks ,to fault tolerant and efficient transmission of information in networks, and to communications between processors in parallel computers.

1) User obtains Ticket Granting Ticket(TGT)

2) Client application uses TGT to request a Service Ticket for Hadoop Service (HDFS/HIVE).
3)Client application connects to Hadoop service providing the service Ticket for authentication.
4)User authenticated using the Service ticket and service key.
5)Results returned from Hadoop service.

3. Advantages:

1.Implementation of network authentication protocol on Hadoop database using a third party application Kerberos.
2. Validation is provided by a trusted third party in the form of the Kerberos Key Distribution Center..
3. Privacy of Kerberos clients is also enhanced. A working implementation of this extended protocol has been developed
4. Security and scalability achieved.

4. CONCLUSIONS

The complexity and cost involved in deploying a Hadoop analytics cluster that might be used only occasionally make the adoption of Hadoop-as-a-Service offerings by public cloud service providers an appealing alternative. However, enterprises that handle sensitive information such as financial and healthcare records have strict requirements regarding security and in particular confidentiality of the data that they move, store, and process in the cloud. Towards easing these security concerns and driving wider adoption of Hadoop cloud services, this work described the architecture and requirements for a Hadoop cloud service that provides end- to-end confidentiality for data-at-rest and allows the enterprise to control access to its data that is stored in the cloud.

The overhead with the introduction of capabilities is low but se- cures the data access for only clients which have been issued the capabilities by Name node**.**

5. ACKNOWLEDGEMENT

We express our sincere thanks to Head of Department of Computer Engineering for her kind co-operation. We express our sincere thanks to Prof. C. S. Mankar.

6. REFERENCES

[1]. SDFS: Secure Distributed File System for Data-at-Rest Security for Hadoop-as a-Service.
Petros Zerfos, Hangu Yeo, Brent D. Paulovicks and Vadim Sheinin .IBM T.J.
Watson Research Center York- town Heights, N Y U.S.A. pzerfos,hangu,ovicks,vadims@us.ibm.com .
[2]. A Token Authentication Solution for Hadoop Based on Kerberos Pre-Authentication .
Kai Zhen, Weihua Jiang Big Data Technologies Intel Corporation Shanghai , China .

[3]. Hadoop with Kerberos Architecture Consideration.
Stuart Rogers, Tom Keefer.

[4].Distributed Authentication in Kerberos Using Public Key Cryptography.