SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)

Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Department of MCA

Topic: Hadoop Security

COURSE	UNIT - III	CLASS
19CAT702 Big Data Analytics	Hadoop Environment	III Semester / II MCA

Session Objectives

- Know the security risks the Big Data analytics
- Understand how security can be handled through Kerberos protocol

Hadoop Security Issues

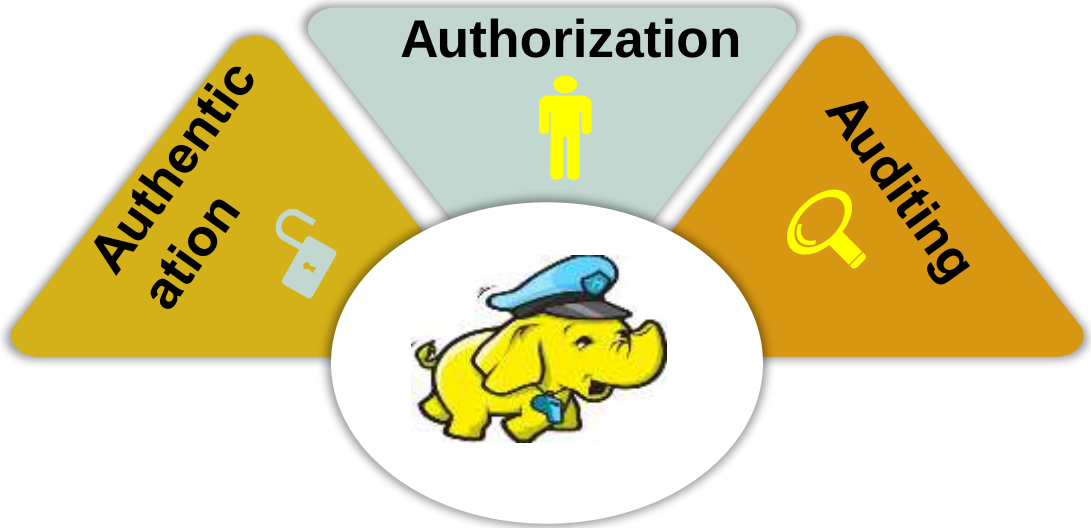
- Insufficient authentication
- No privacy and integration
  - Insecure network transport
  - No message level security
- Arbitrary code execution
  - Malicious user can submit a job

Security Risks

- HDFS provides authentication mechanism only
- Sensitive data is accessible to a small set of user
- Less sensitive data may be made available to larger set of users
- Data protection, secure authentication must be in place for shared clusters

Hadoop Security

Procedure to secure data storage unit Hadoop by offering well defined security layer against cyber threat



Hadoop Security

Kerberos, a mature open-source network authentication protocol, to authenticate the user

Kerberos

- Hadoop is configurable in either secure or non-secure mode
- Kerberos is the basis for authentication in Hadoop secure mode
- Network Authentication Protocol to provide powerful authentication services to both Server and Client ends through Secret key cryptography techniques
- It usses encrypted service tickets throughout the entire session
- Kerberos says that a user is who they say they are
- Hadoop’s job to determine whether that user has permission to perform a given action

Kerberos

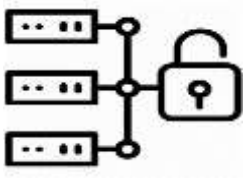
- Design requirements
  - Interaction between server and client encrypted
  - Based on Secret key distribution model
  - Convenient for user
  - Protect against intercepted credentials

Kerberos - Entities in Workflow

Client: Acts on behalf of the user to initiate service request

Server: Host the Services

Authentication server: Performs client authentication



Key Distribution Center: Authentication Server+Ticket Grant Server

Ticket Grant Server: Application server issues ticket as a service

Kerberos

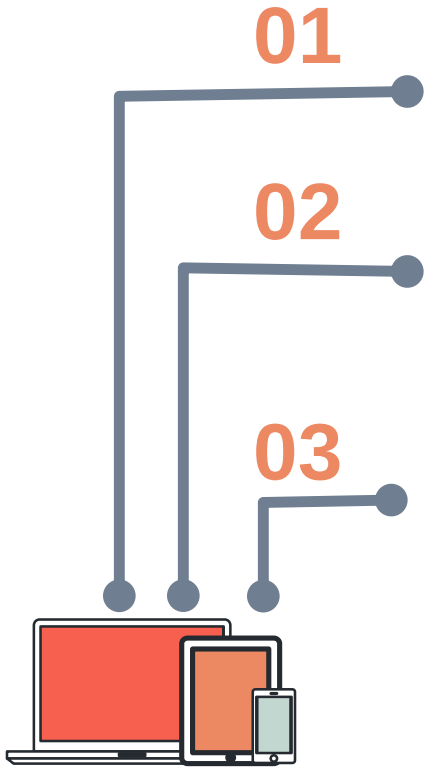
Three steps that a client must take to access a service when using Kerberos, each of which involves a message exchange with a server

To enable Kerberos, configure the core-site.xml as follows

- *hadoop.security.authorization=true*
- *hadoop.security.authentication=Kerberos (the default is "simple")*

Kerberos

Kerberos Procedure

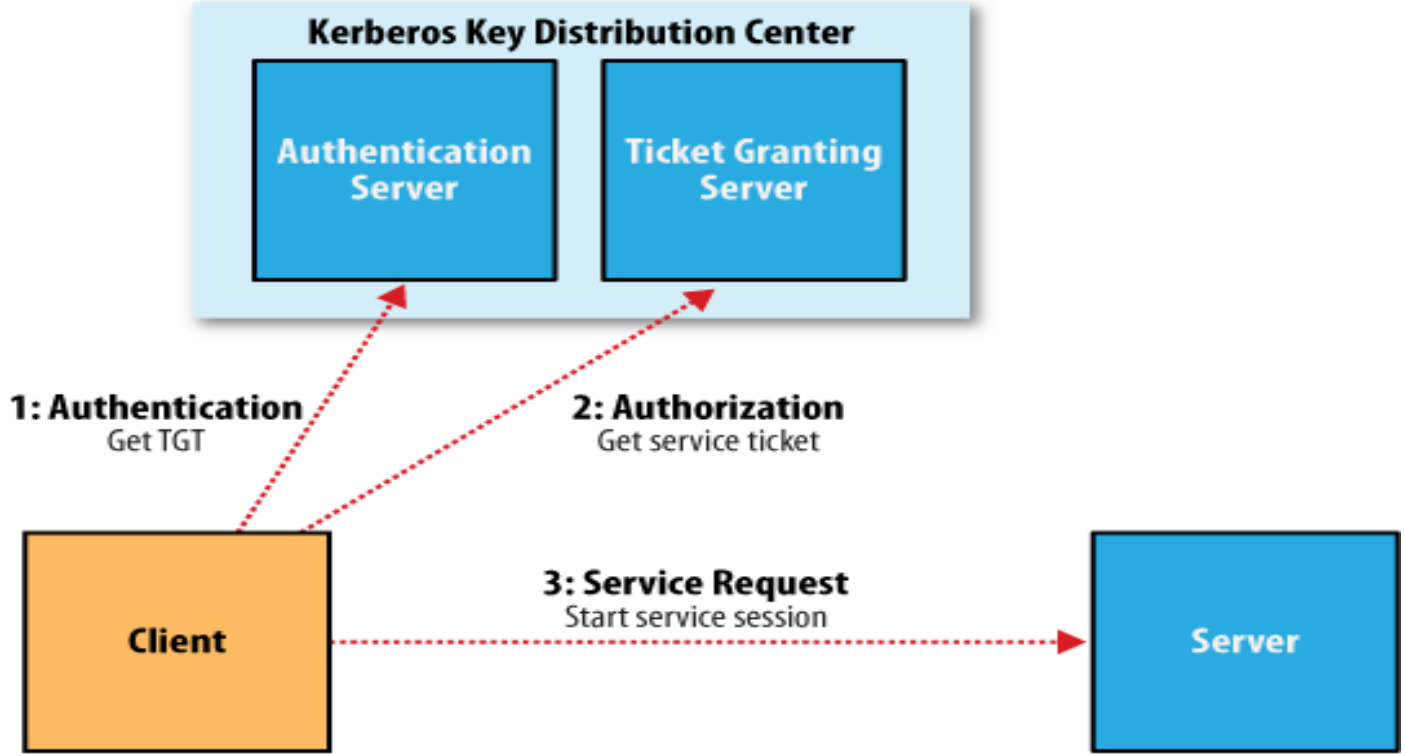


Authentication The client authenticates itself to the Authentication Server and receives a time stamped Ticket-Granting Ticket (TGT)

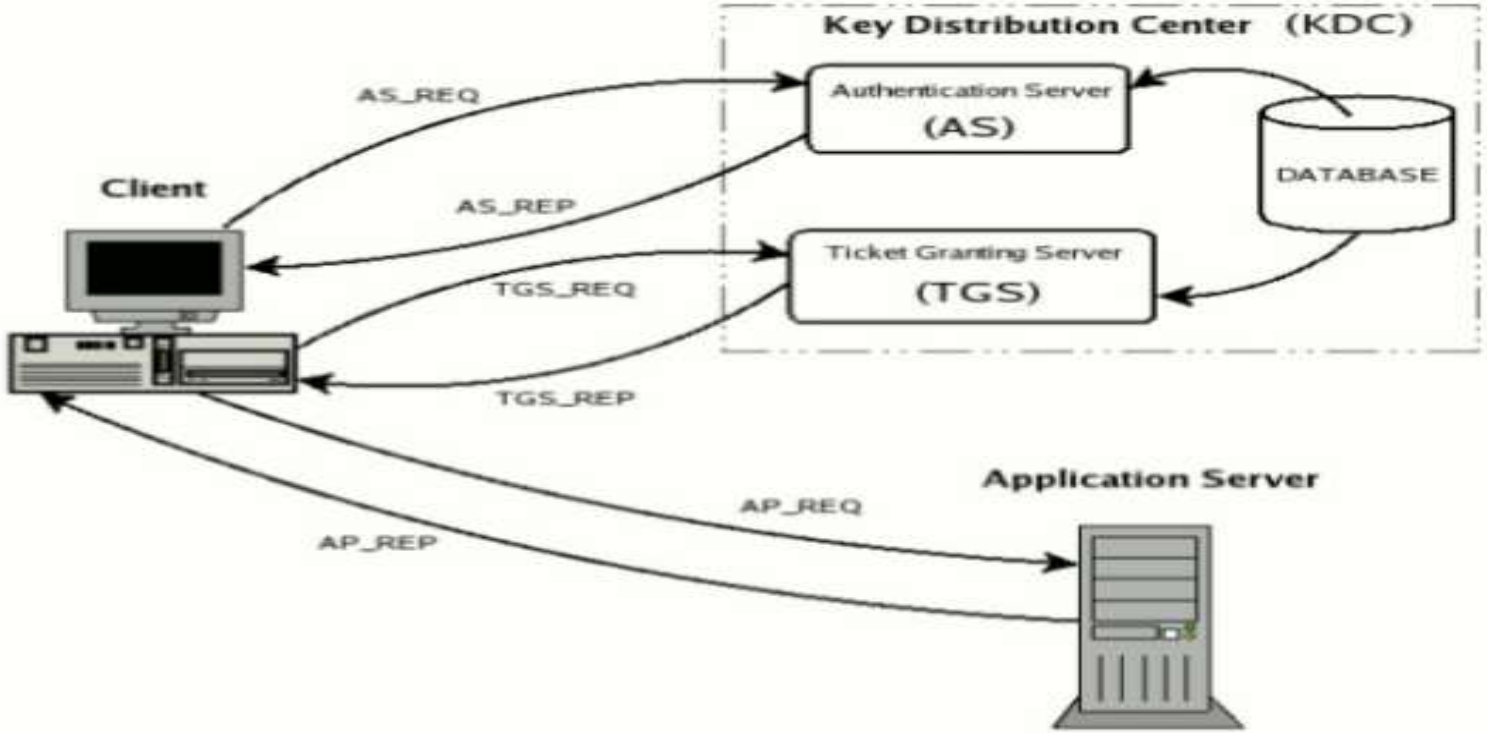
Authorization The client uses the TGT to request a service ticket from the Ticket Granting Server

Service Request The client uses the service ticket to authenticate itself to the server that is providing the service the client is using. In the case of Hadoop, this might be the namenode or the jobtracker

Kerberos Procedure



Kerberos


Key Distribution Center (KDC)
Authentication Server
(AS)
DATABASE
Ticket Granting Server
(TGS)
Client
AS_REQ
AS_REP
TGS_REQ
TGS_REP
Application Server
AP_REQ
AP_REP

Delegation Token

- In a distributed system, many client-server interactions, each of which must be authenticated
- To avoid high load on the KDC on a busy cluster, Hadoop uses delegation tokens to authenticated access without having to contact the KDC again
- Token is generated by the server(namenode) and can be thought of as a shared secret between the client and server

Delegation Token

- On the first RPC call to the namenode, the client has no delegation token, so it uses Kerberos to authenticate, and as a part of the response it gets a delegation token from the namenode
- Delegation tokens are used by the jobtracker and tasktrackers to access HDFS during course of the job
- When the job has finished, the delegation tokens are invalidated
- Token has an expiration time and require periodic renewals to keep their validity.

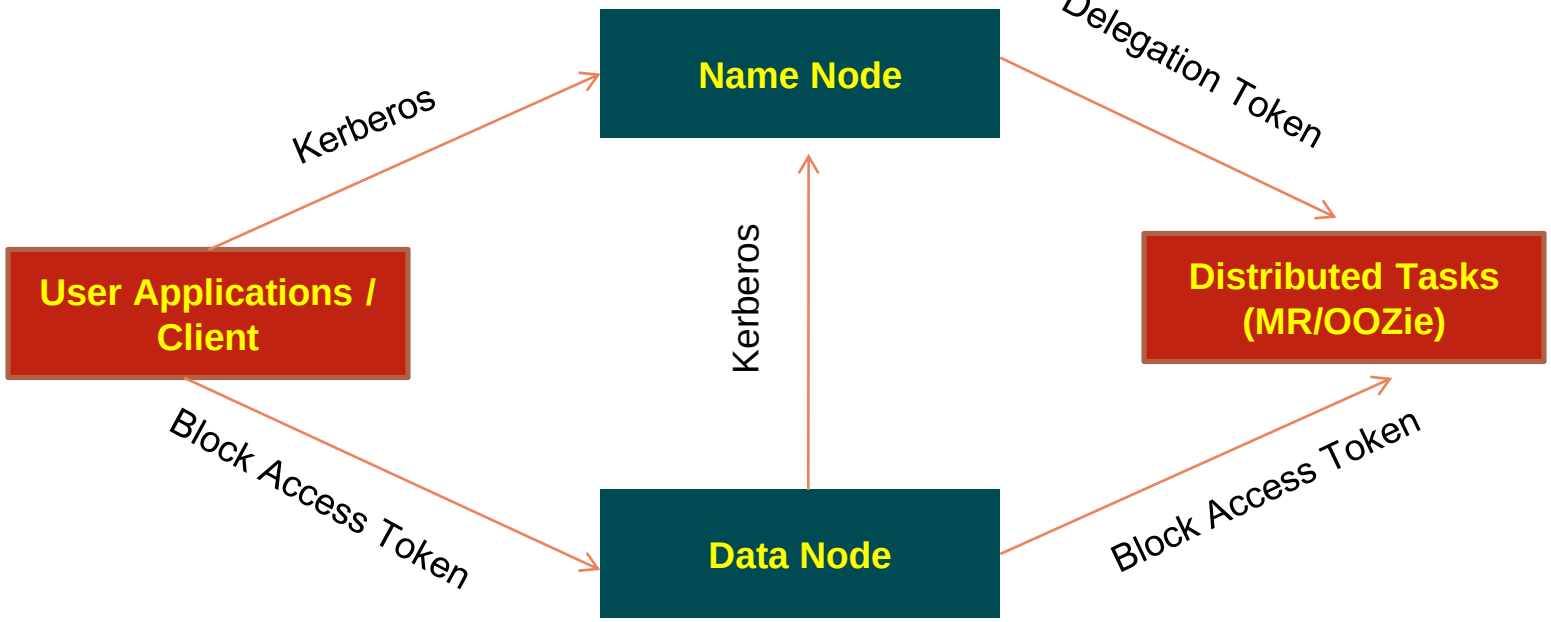
Block Access Token

- To perform operations on HDFS blocks, client uses block access token
- Namenode passes to the client in response to a metadata request
- Client uses the block access token to authenticate itself to datanodes
- Namenode shares its secret key used to generate the block access token with datanodes

Other Security Enhancements

- Tasks can be run using the operating system account for the user who submitted the job, so, OS is used to isolate running tasks, so they can't send signals to each other
- Files that are world-readable are put in a shared cache (the insecure default), otherwise they go in a private cache
- Users can view and modify only their own jobs, not others

HDFS Authentication



Other Security Enhancements

The shuffle is secure, preventing a malicious user from requesting another user's map outputs

A datanode may be run on a privileged port

A task may only communicate with its parent tasktracker, thus preventing an attacker from obtaining MapReduce data from another user's job

Session Objectives

- **Kerberos protocol is used for**

  A. Authorization B. Authentication C.User Mgt D. Communication

- **Ticket-Granting Ticket is generated by**

  A. Authorization server B. Ticket Granting Server C. Granting server

- **Key Distribution Center is a combination of**

  A. Authorization server + Ticket Granting Server
  B. Authentication server + Ticket Granting Server
  C. Client + Ticket Granting Server
  D. Server + Ticket Granting Server

References

- Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 4th Edition, 2012

Web Resources

- https://freevideolectures.com/course/3610/hadoop-administration/12
- https://www.wideskills.com/hadoop/hadoop-administration
- https://www.tutorialspoint.com/map_reduce/map_reduce_hadoop_administration.htm


Thank u