Hadoop Security

Hadoop Security provides authentication, authorization, auditing, and secure the Hadoop data storage unit by offering an inviolable wall of security against any cyber threat.

- Hadoop uses Kerberos to authenticate the user accessing the Hadoop HDFS files or directories.
- Transparent encryption is used for protecting the files or directories in HDFS.
- HDFS checks for the client permission to access the files or directories.

Hadoop with Kerberos

1. Authentication**:** In Kerberos, the client first authenticates itself to the authentication server. The authentication server provides the timestamped Ticket-Granting Ticket (TGT) to the client.
2. Authorization: The client then uses TGT to request a service ticket from the Ticket-Granting Server.
3. Service Request: On receiving the service ticket, the client directly interacts with the Hadoop cluster daemons such as NameNode and ResourceManager.

Transparent Encryption in HDFS

1. For data protection, Hadoop HDFS implements transparent encryption.
2. Once it is configured, the data that is to be read from and written to the special HDFS directories is encrypted and decrypted transparently without requiring any changes to the user application code.

HDFS file with directory permission

1. For authorizing the user, the Hadoop HDFS checks the files and directory permission after the user authentication.
2. Every file and directory in HDFS are having an owner and a group.
3. The files or directories have different permissions for the owner, group members, and all other users.
4. For files, r is for reading permission, w is for write or append permission.
5. For directories, r is the permission to list the content of the directory, w is the permission to create or delete files/directories, and x is the permission to access a child of the directory.

Tools used for Hadoop Security

The two major Apache open-source projects that support Hadoop Security are Knox and Ranger.

1. Knox

   Knox is a REST API base perimeter security gateway that performs authentication, support monitoring, auditing, authorization management, and policy enforcement on Hadoop clusters. It allows only the successfully authenticated users to access the Hadoop cluster.

2. Ranger

   It is an authorization system that provides or denies access to Hadoop cluster resources such as HDFS files, Hive tables, etc. based on predefined policies. User request assumes to be already authenticated while coming to Ranger. It has different authorization functionality for different Hadoop components such as YARN, Hive, HBase, etc.