Introduction to Hadoop Security

Today, data explosion is a reality of the digital universe and the amount of data extremely increases even every second. Hadoop has its heart in storing and processing large amounts of data efficiently and as it turns out, cheaply when compared to other platforms.

As the adoption of Hadoop increases, the volume of data and the types of data handled by Hadoop deployments have also grown. For production deployments, a large amount of this data is generally sensitive or subject to industry regulations and governance controls.

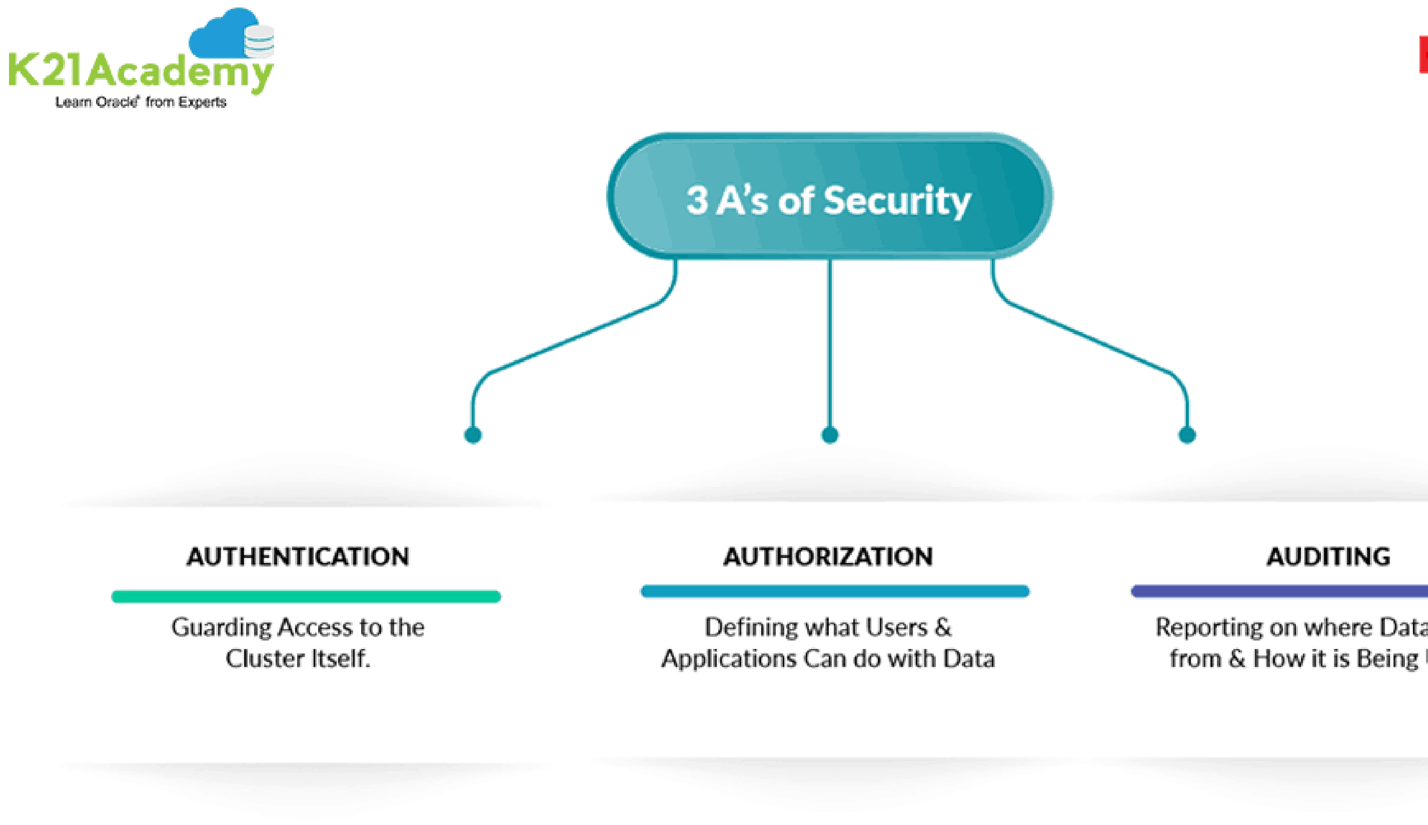
In order to be compliant with such regulations, Hadoop must offer strong capabilities to thwart any attacks on the data it stores, and take measures to ensure proper security at all times. The security scene for Hadoop is evolving rapidly. However, this rate of progress isn't steady overall Hadoop parts, which is the reason the level of security capacity may seem uneven over the Hadoop environment. That is, a few segments may be perfect with more grounded security innovations than others.

Why Hadoop Security Is Important?

- **Laws governing data privacy:** Particularly important for healthcare and finance industries
- Export control regulations for defense information
- Protection of proprietary research data
- Company policies
- Di erent teams in a company have di erent needs
- **Setting up multiple clusters is a common solution:** One cluster may contain protected data, another cluster does not

The Three A's of Security and Data Protection

So what to do about Hadoop security? In part, security and governance in Hadoop require many of the same approaches seen in the traditional data management world. These include the "Three As" of security and data protection.


K21Academy
Learn Oracle from Experts
3 A's of Security
AUTHENTICATION
Guarding Access to the Cluster Itself.
AUTHORIZATION
Defining what Users & Applications Can do with Data
AUDITING
Reporting on where Data from & How it is Being

Authorization: It is the process of determining what data, types of data or applications that user is allowed to access

- Determining whether a participant is allowed to perform an action
- Typically done by checking an access control list

Authentication: It is simply the process of accurately determining the identity of a given user attempting to access a

Hadoop cluster or application based on one of a number of factors.

- Confirming the identity of a participant
- Typically done by checking credentials (username/password)

Auditing: It is the process of recording and reporting what an authenticated, authorized user did once granted access to the cluster, including what data was accessed/changed/added and what analysis occurred.

Data Protection: Data protection refers to the use of techniques such as encryption and data masking to prevent sensitive data from being accessed by unauthorized users and applications.

Types of Hadoop Security

- HDFS file ownership and permissions
- Enhanced security with Kerberos
- Encrypted HDFS data transfers
- HDFS data at rest encryption
- Encrypted HTTP tra c

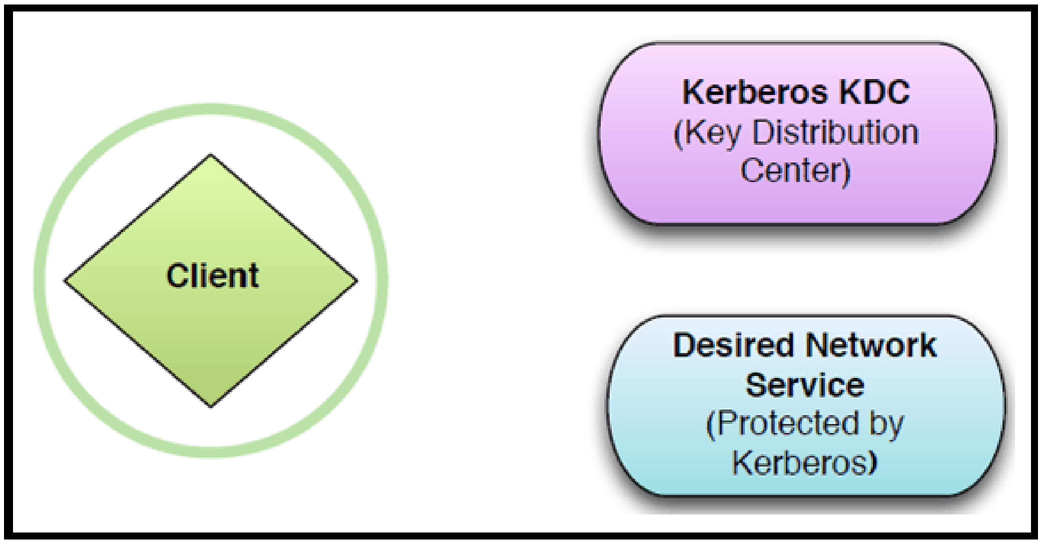
What Kerberos is and How it Works?

The Role of Kerberos in CDH5

- Hadoop daemons leverage Kerberos to perform user authentication on all remote procedure calls (RPCs)
- Group resolution is performed on master nodes to guarantee group membership is not manipulated

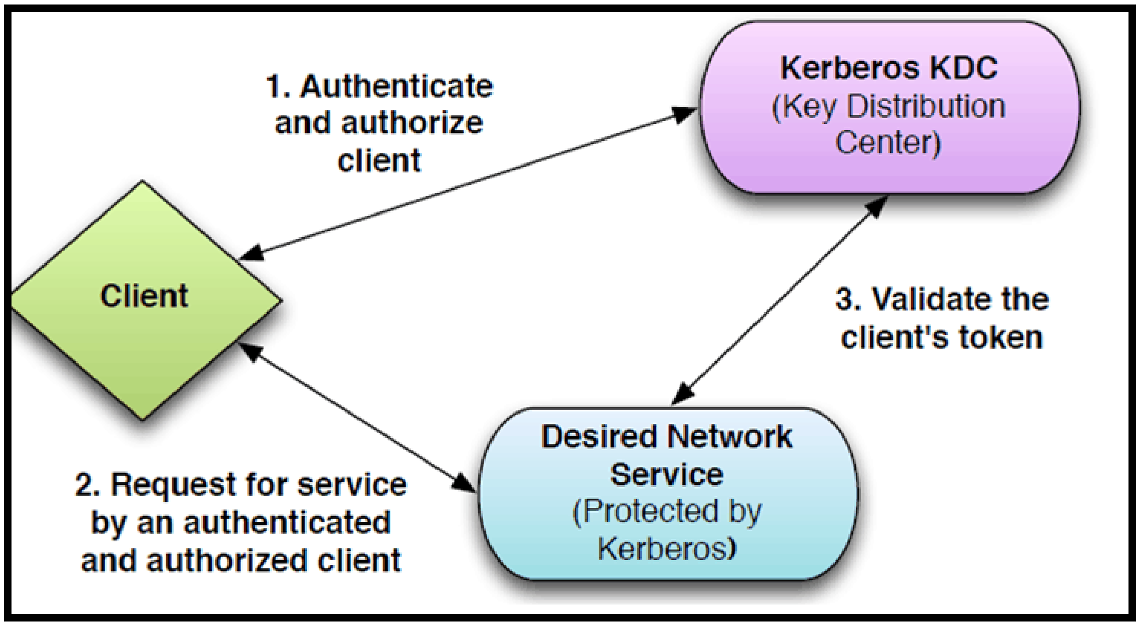
Kerberos Exchange Participants

- Kerberos involves messages exchanged among three parties
    - The client
    - The server providing the desired network service
    - The Kerberos Key Distribution Center (KDC)


Client
Kerberos KDC (Key Distribution Center)
Desired Network Service (Protected by Kerberos)

General Kerberos Concepts

- Kerberos is a standard network security protocol
    - Currently at version 5 (RFC 4120)
    - Services protected by Kerberos don’t directly authenticate the client


1. Authenticate and authorize client
Kerberos KDC (Key Distribution Center)
Client
3. Validate the client's token
Desired Network Service (Protected by Kerberos)
2. Request for service by an authenticated and authorized client

Essential Points

- Kerberos is the primary technology for enabling authentication security on the cluster
  - Manual configuration requires many steps
  - We recommend using Cloudera Manager to enable Kerberos
- Encryption can be enabled at the filesystem level, HDFS level, and the network level
- Sentry enables security for Hive and Impala

NNThroughputBenchmark, as its name indicates, is a name-node throughput benchmark, which runs a series of client threads on a single node against a name-node. If no name-node is configured, it will firstly start a name-node in the same process (*standalone mode*), in which case each client repetitively performs the same operation by directly calling the respective name-node methods. Otherwise, the benchmark will perform the operations against a remote name-node via client protocol RPCs (*remote mode*). Either way, all clients are running locally in a single process rather than remotely across different nodes. The reason is to avoid communication overhead caused by RPC connections and serialization, and thus reveal the upper bound of pure name-node performance.

The benchmark first generates inputs for each thread so that the input generation overhead does not effect the resulting statistics. The number of operations performed by threads is practically the same. Precisely, the difference between the number of operations performed by any two threads does not exceed 1. Then the benchmark executes the specified number of operations using the specified number of threads and outputs the resulting stats by measuring the number of operations performed by the name-node per second.