BIG DATA SECURITY IN HADOOP

201806100110091,201806100110127

What is hadoop ?

Hadoop is

an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly

Why do we need Hadoop Security?

Hadoop is the most powerful, robust and highly scalable big data processing framework capable enough to crunch petabytes of data with ease. Due to its unmatched capabilities, back then, every business sector, health, military and finance departments started using Hadoop

Hadoop started gaining popularity. This is when the Hadoop developers found a monumental miscalculation. *Hadoop lacked a dedicated security software* inside it. This affected many ares.

Why do we need Hadoop Security?

1

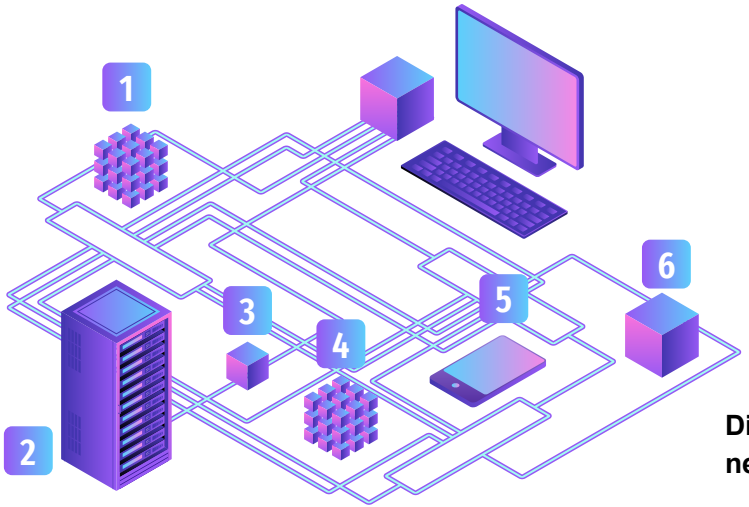
Laws governing data privacy: Particularly important for healthcare and finance industries

2

Export control regulations for defense information

3

Protection of proprietary research data



4

Company policies

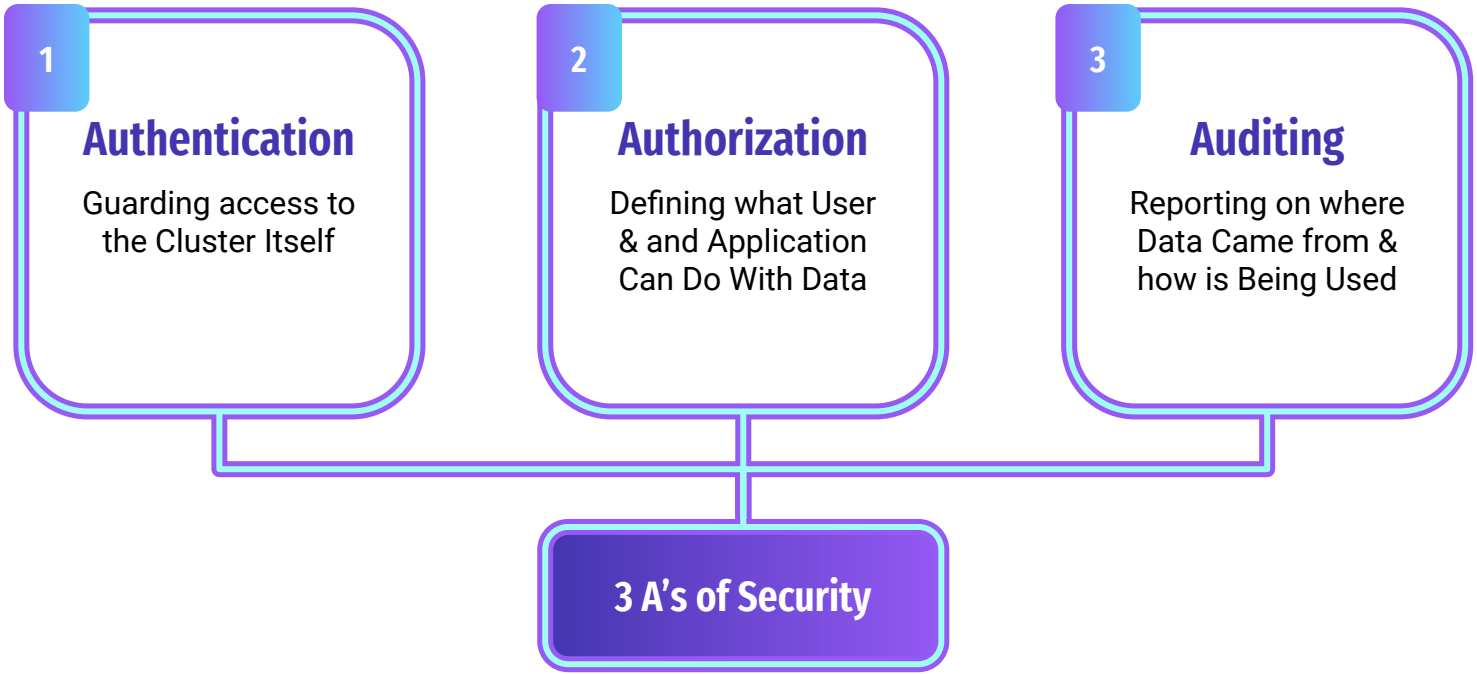
5

Different teams in a company have different needs

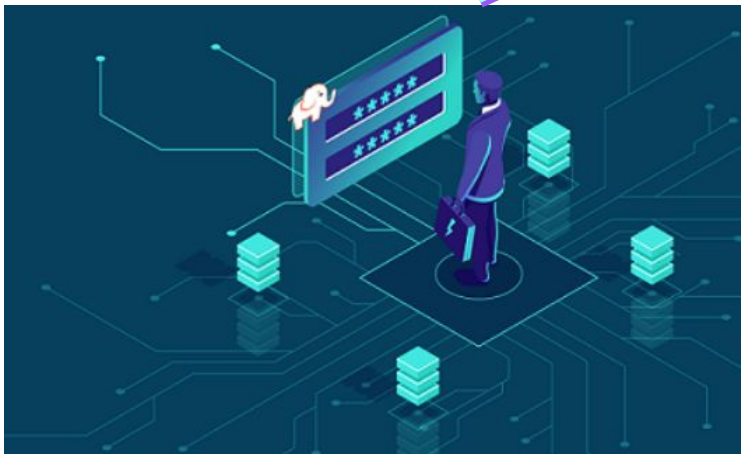
6

Setting up multiple clusters is a common solution: One cluster may contain protected data, another cluster does not

The Three A's of Security & Data Protection



Authentication

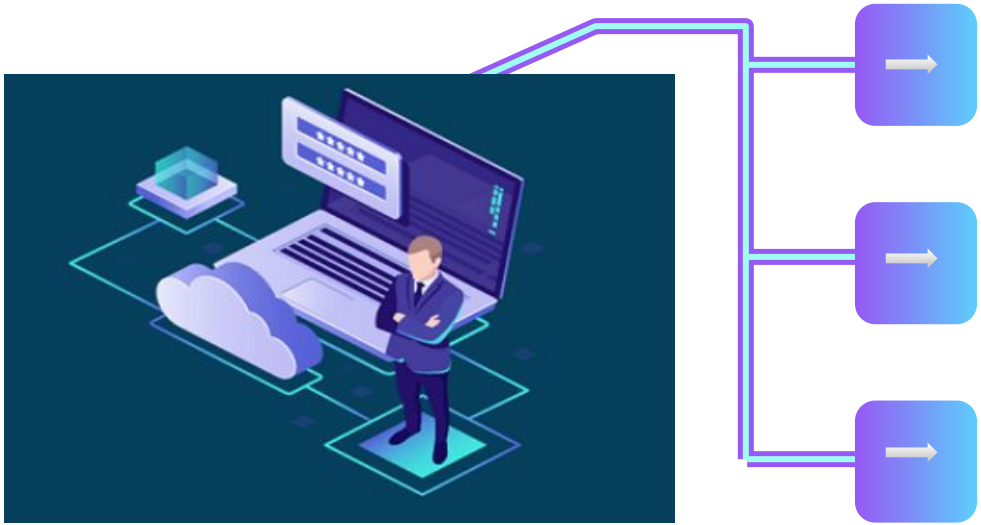


It is the process of determining what data, types of data or applications that user is allowed to access

Determining whether a participant is allowed to perform an action

Typically done by checking an access control list

Authorization

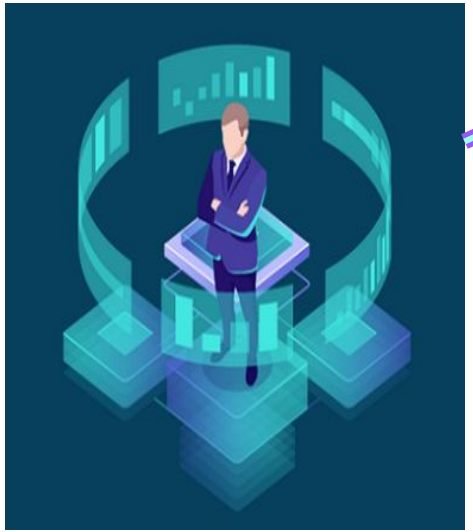


It is simply the process of accurately determining the identity of a given user attempting to access a Hadoop cluster or application based on one of a number of factors.

Confirming the identity of a participant

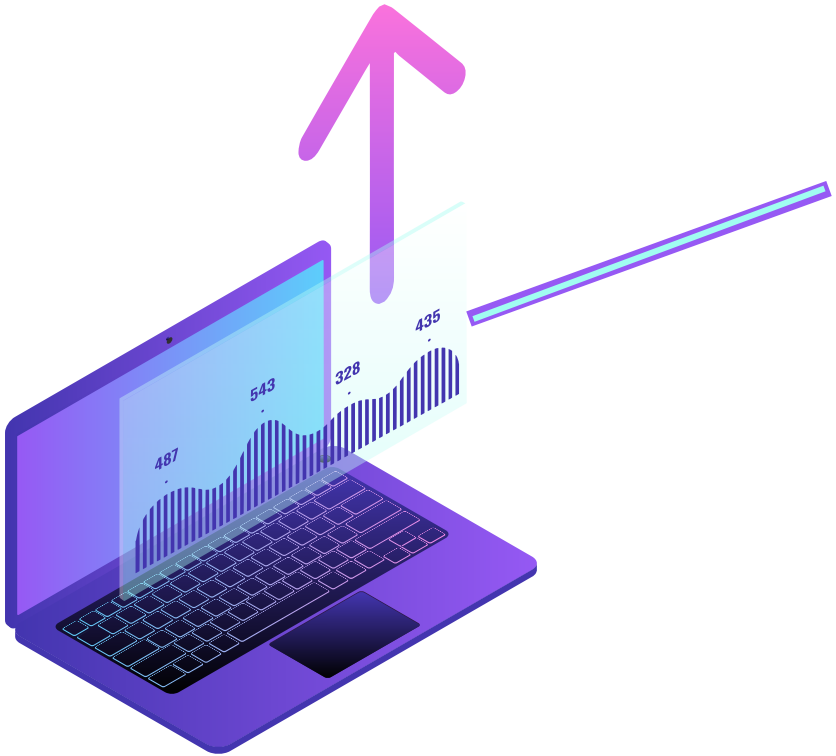
Typically done by checking credentials (username/password)

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

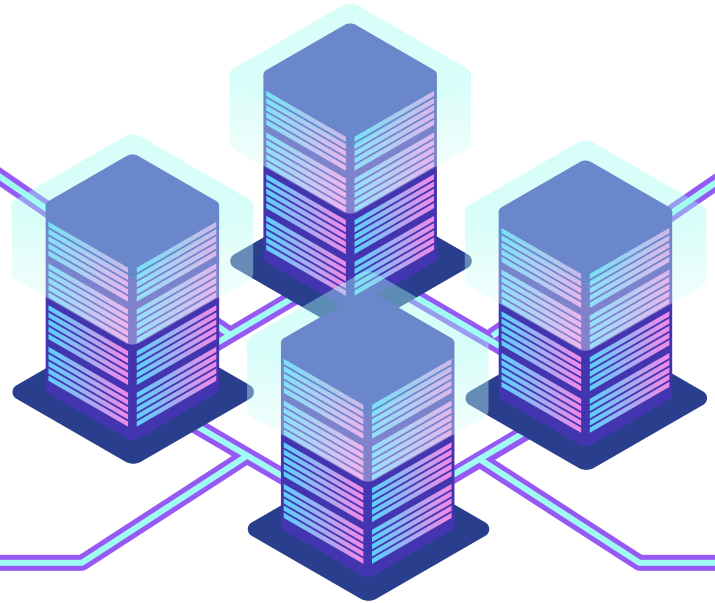


Data protection refers to the use of techniques such as encryption and data masking to prevent sensitive data from being accessed by unauthorized users and applications.

Types of Hadoop Security

1

Enhanced security with Kerberos



2

Encrypted HDFS data transfers

4

HDFS file ownership and permissions

3

HDFS data at rest encryption Encrypted HTTP traffic

kerberos Security

Kerberos is one of the leading Network Authentication Protocol designed to provide powerful authentication services to both Server and Client-ends through Secret-Key cryptography techniques. It is proven to be highly secure since it uses encrypted service tickets throughout the entire session.

HDFC Encryption

HDFS Encryption is a formidable advancement that Hadoop ever embraced. Here, the data from source to destination(HDFS) gets completely encrypted. This procedure does not require any changes to be made on to the original Hadoop Application, making the client to be the only authorized personnel to access the data.

Traffic Encryption

Traffic Encryption is none other than HTTPS(HyperText Transfer Protocol Secure). This procedure is used to secure the data transmission, from the website as well as data transmission to the website. Much online banking gateways use this method to secure transactions over a Security Certificate

HDFS File and Directory Permission

HDFS file directory permissions work in a simple POSIX format. The Read and Write permissions are provided as r and s respectively. The permissions to the Super User and Client are set differently based on the confidentiality of the file.