

SECURING BIG DATA IN HADOOP

WORKSHOP

HTTPS://GITHUB.COM/CHICOLINUX/DC26

BY MIKE GUIRAO
@miguelguirao

SO… WHAT IS APACHE HADOOP?



The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

APACHE HADOOP

The project includes these modules:

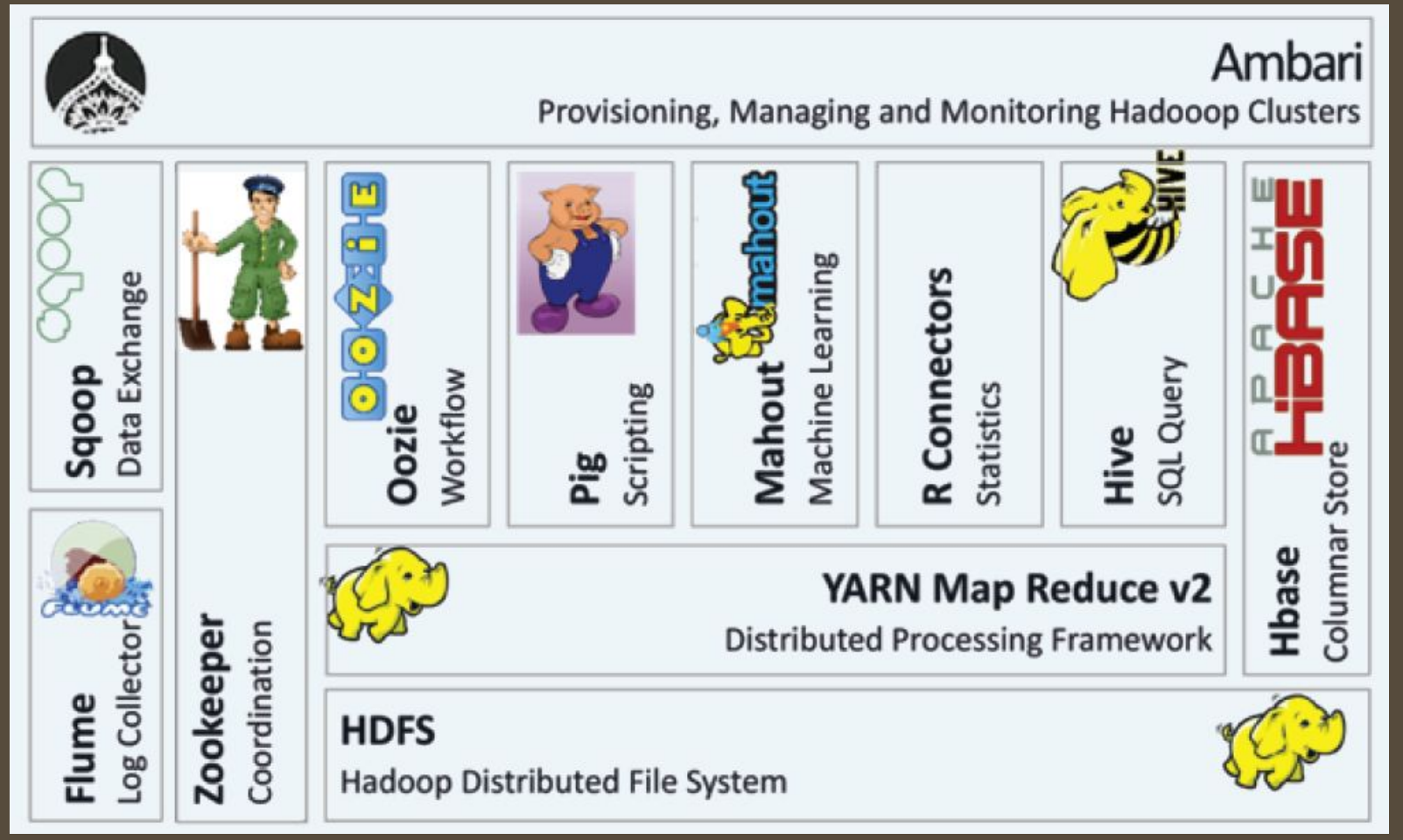
Hadoop Common: The common utilities that support the other Hadoop modules.

Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.

Hadoop YARN: A framework for job scheduling and cluster resource management.

Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

APACHE HADOOP ECOSYSTEM





ARE YOU ALL SETUP AND READY TO GO?

30 MIN

APACHE AMBARI

The Apache Ambari project is aimed at making Hadoop management simpler by developing software for:

- provisioning,
- managing, and
- monitoring Apache Hadoop clusters.



Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.



AMBARI LAB

30 MIN

APACHE KNOX

The Apache Knox™ Gateway is an Application Gateway for interacting with the REST APIs and UIs of Apache Hadoop deployments.

The Knox Gateway provides a single access point for all REST and HTTP interactions with Apache Hadoop clusters.



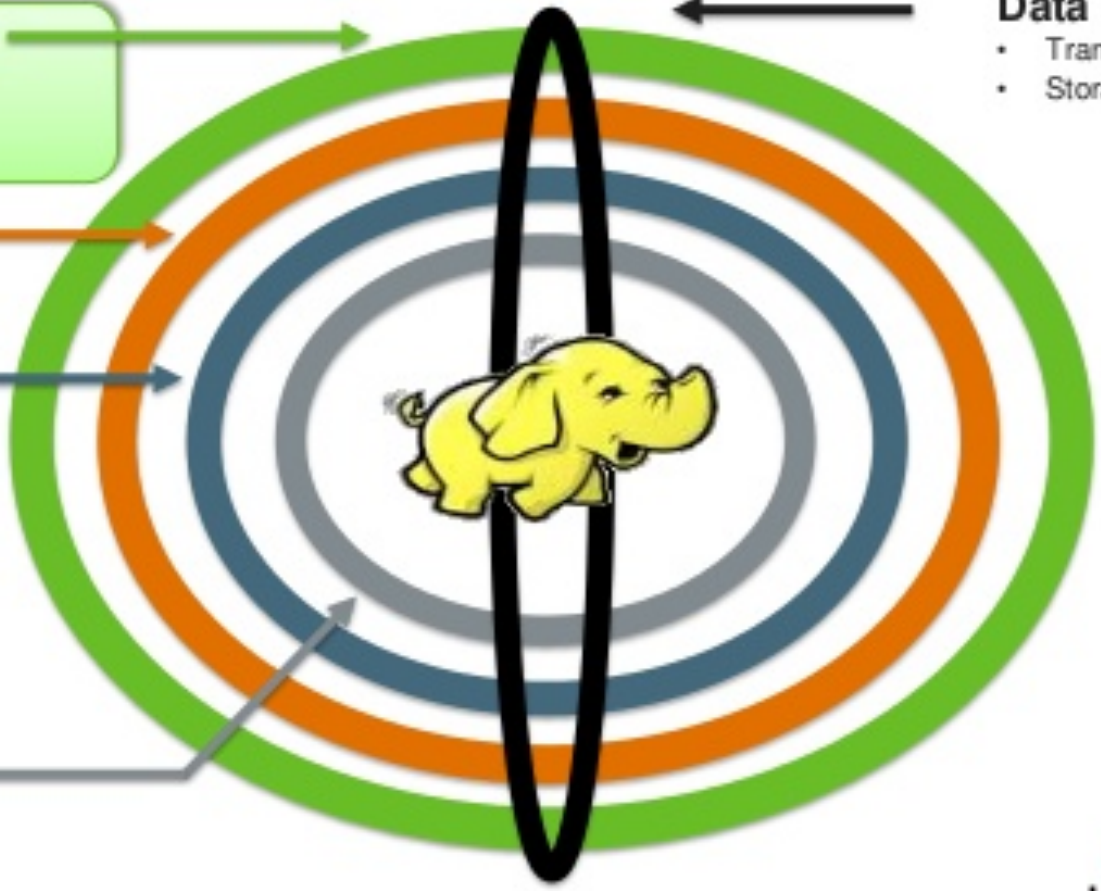
Layers Of Hadoop Security

Perimeter Level Security
- Network Security (i.e. Firewalls)
- Apache Knox (i.e. Gateways)

Authentication
- Kerberos
- Delegation Tokens

Authorization
- MR ACLs
- HDFS Permissions
- HDFS ACLs
- HiveATZ-NG
- HBase ACLs
- Accumulo Label Security
- XA Security Policies

OS Security
- File Permissions
- Process Isolation

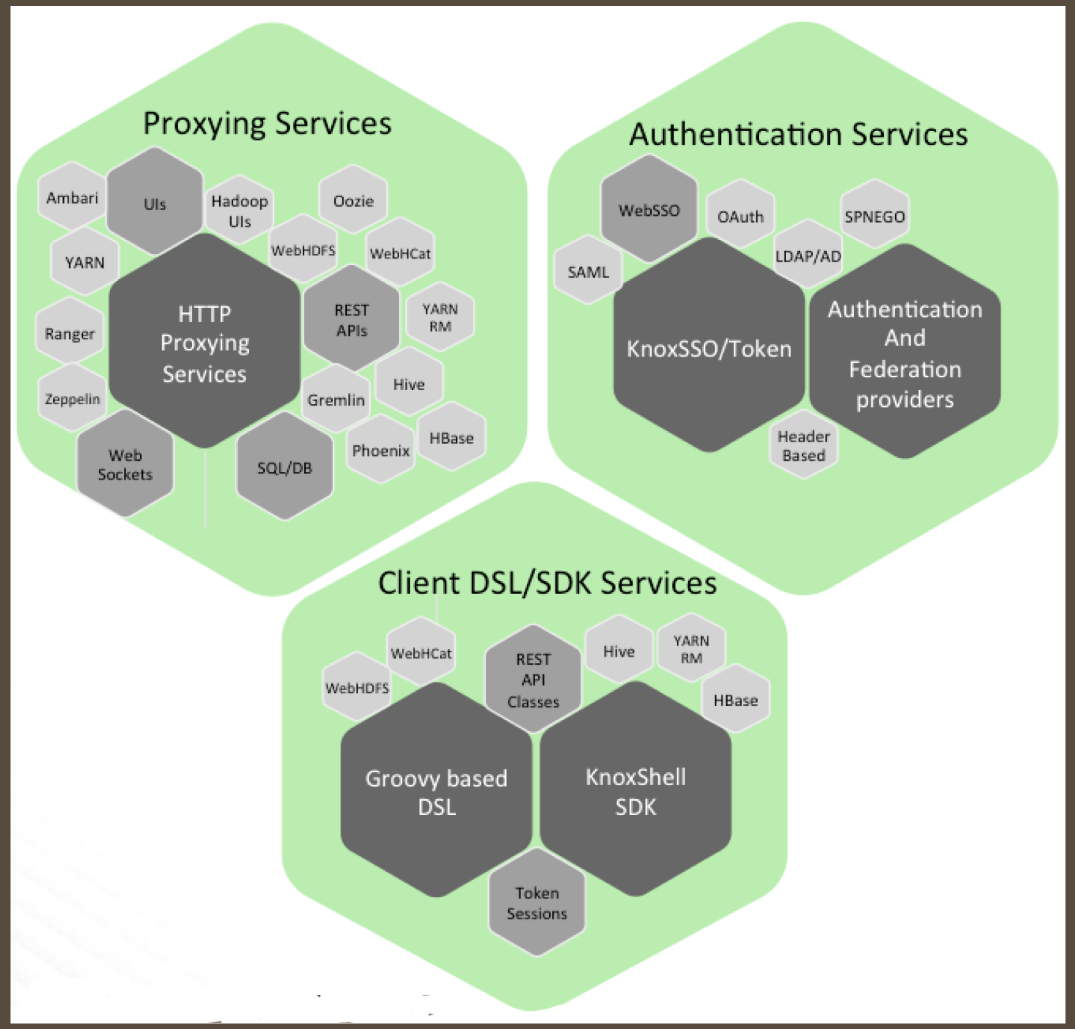


Data Protection
- Transport
- Storage

© Hortonworks Inc. 2014

Hortonworks

KNOX DELIVERS THREE GROUPS OF USER FACING SERVICES:



SUPPORTED APACHE HADOOP SERVICES

1) Ambari
2) WebHDFS (HDFS)
3) Yarn RM
4) Stargate (Apache HBase)
5) Apache Oozie
6) Apache Hive/JDBC
7) Apache Hive WebHCat (Templeton)
8) Apache Storm

1) Apache Tinkerpop - Gremlin
2) Apache Avatica/Phoenix
3) Apache SOLR
4) Apache Livy (Spark REST Service)
5) Kafka REST Proxy

SUPPORTED APACHE HADOOP ECOSYSTEM UIS

1) Name Node UI
2) Job History UI
3) Yarn UI
4) Apache Oozie UI
5) Apache HBase UI

1) Apache Spark UI
2) Apache Ambari UI
3) Apache Ranger Admin Console
4) Apache Zeppelin
5) Apache NiFi

GOALS OF THE GATEWAY:

- Provide perimeter security for Hadoop REST APIs to make Hadoop security easier to setup and use
  - Provide authentication and token verification at the perimeter
  - Enable authentication integration with enterprise and cloud identity management systems
  - Provide service level authorization at the perimeter
- Expose a single URL hierarchy that aggregates REST APIs of a Hadoop cluster
  - Limit the network endpoints (and therefore firewall holes) required to access a Hadoop cluster
  - Hide the internal Hadoop cluster topology from potential attackers



APACHE KNOX LAB

60 MIN

APACHE RANGER

Apache Ranger™ is a framework to enable, monitor and manage comprehensive data security across the Hadoop platform.

The vision with Ranger is to provide comprehensive security across the Apache Hadoop ecosystem.



APACHE RANGER GOALS OVERVIEW

- Centralized security administration to manage all security related tasks in a central UI or using REST APIs.
- Fine grained authorization to do a specific action and/or operation with Hadoop component/tool and managed through a central administration tool
- Standardize authorization method across all Hadoop components.
- Enhanced support for different authorization methods - Role based access control, attribute based access control etc.
- Centralize auditing of user access and administrative actions (security related) within all the components of Hadoop.



APACHE RANGER LAB

60 MIN

WHAT IS KERBEROS?



- Kerberos is a network authentication protocol. It is designed to provide strong authentication for client/server applications by using secret-key cryptography.
- The Kerberos protocol uses strong cryptography so that a client can prove its identity to a server (and vice versa) across an insecure network connection.
- After a client and server has used Kerberos to prove their identity, they can also encrypt all of their communications to assure privacy and data integrity as they go about their business.

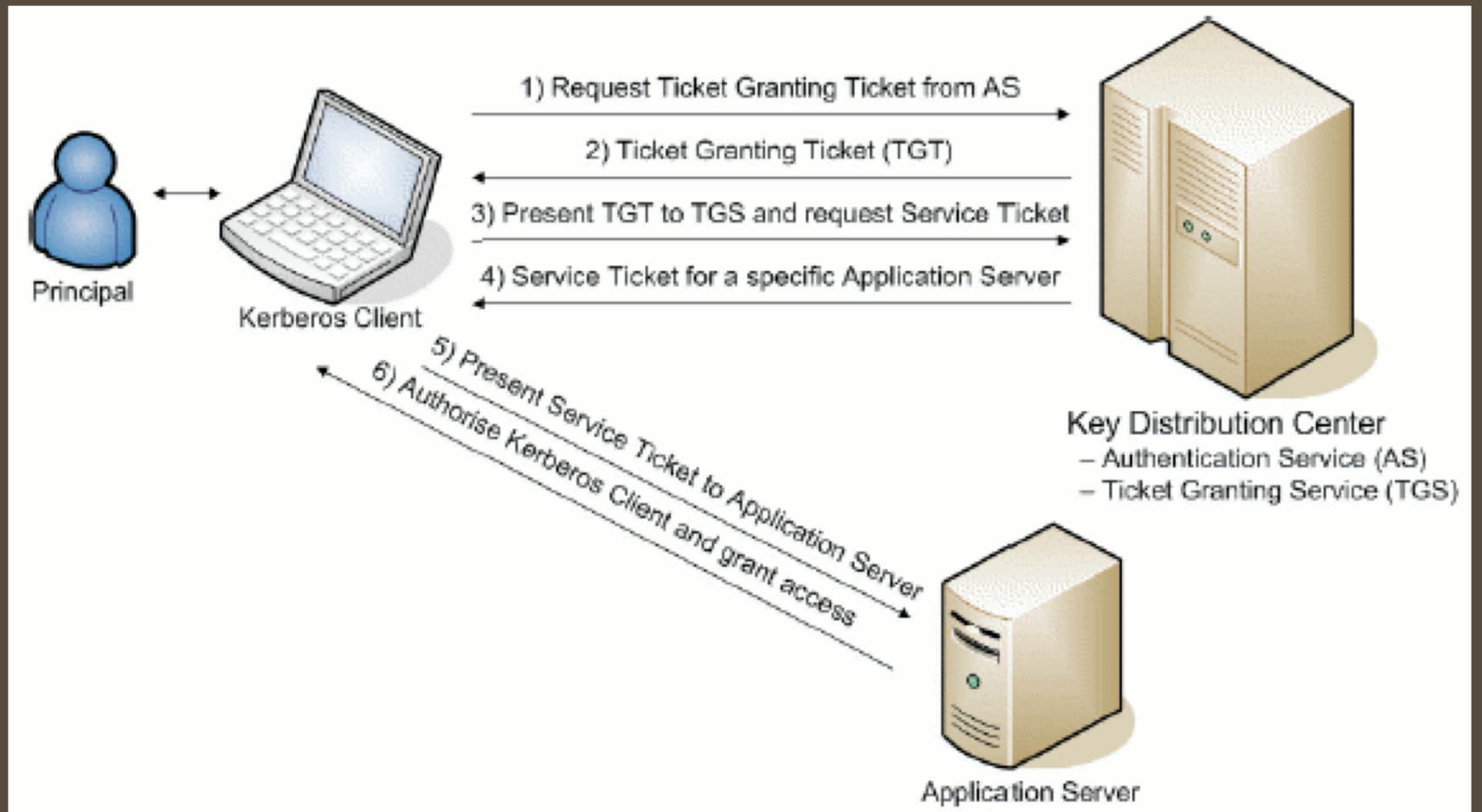


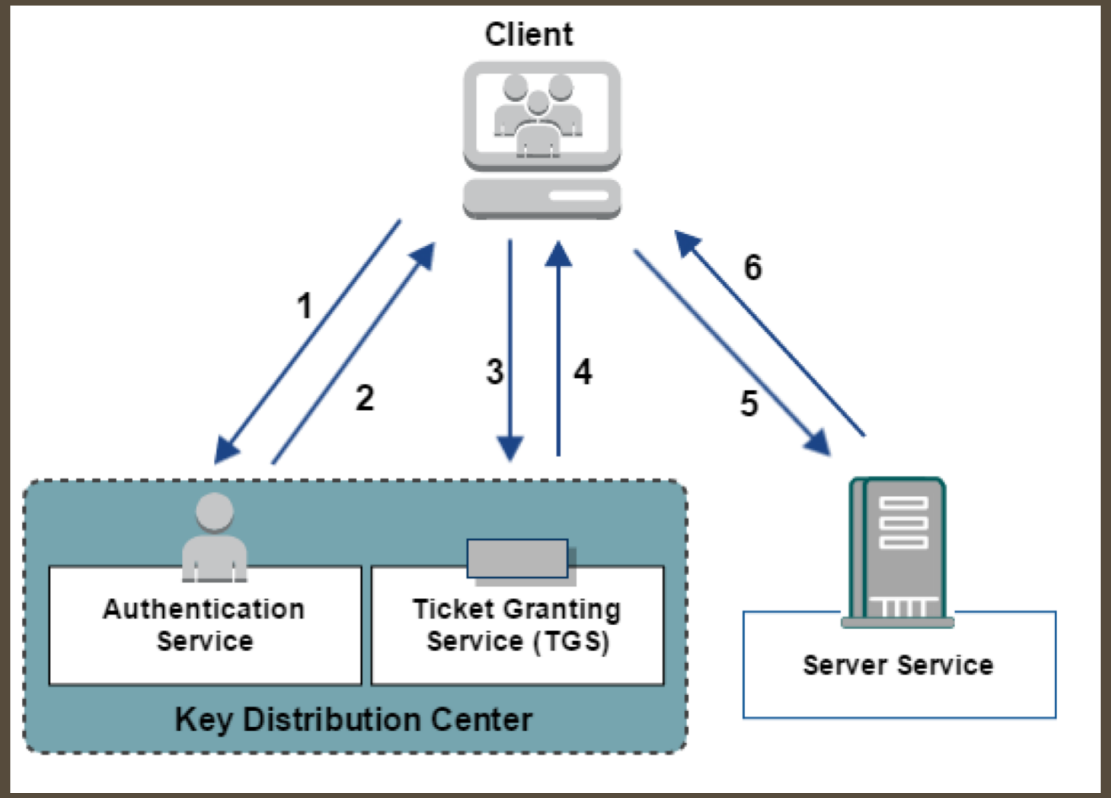
Figure 1: Java Kerberos System Architecture

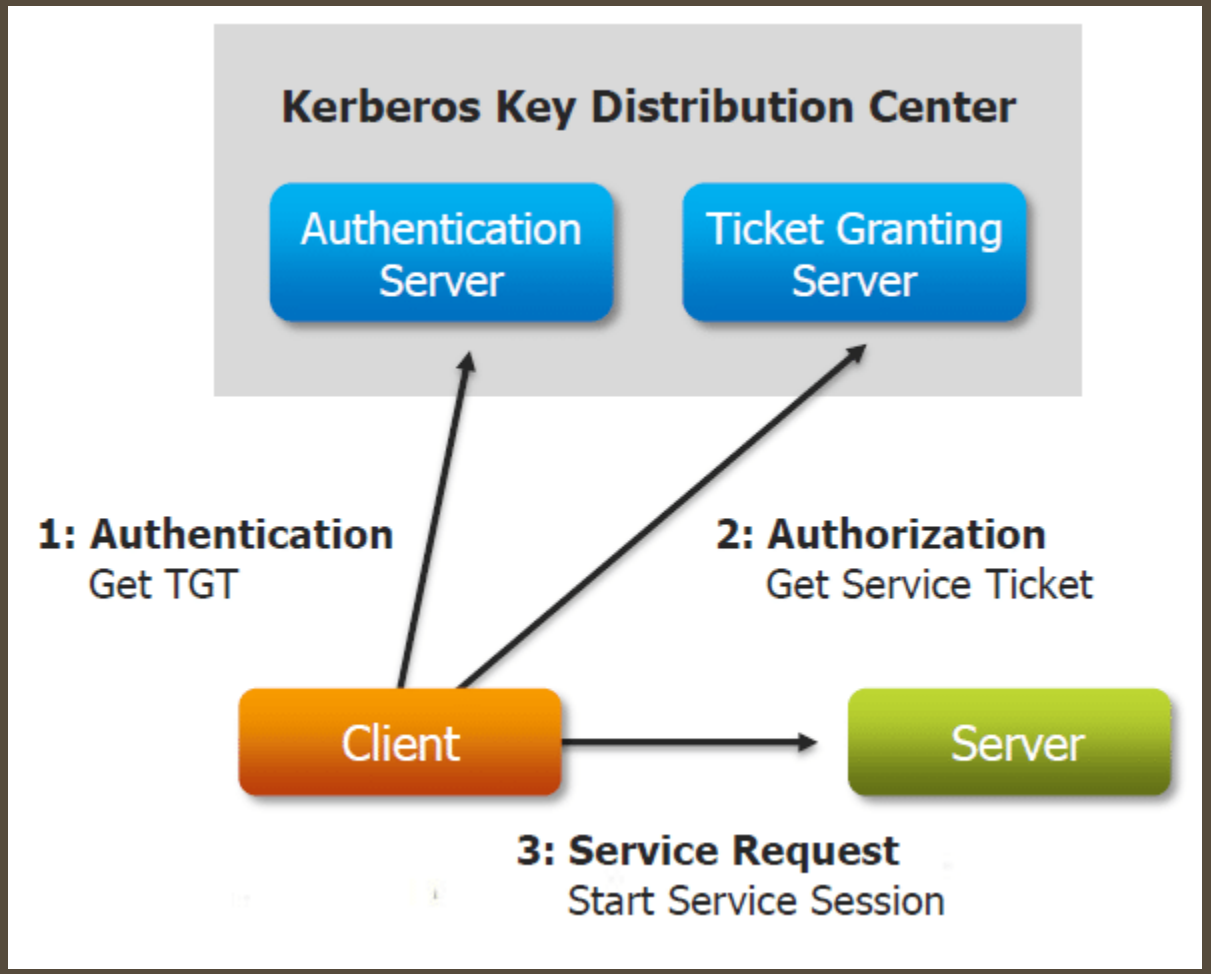
HOW DOES IT WORK?

KERBEROS RELATED TERMS

- PRINCIPAL. A server or client to which Kerberos can assign tickets.
- AUTHENTICATION SERVER (AS). A server that gives authorizes the principal and connects them to the Ticket Granting Server (TGS).
- TICKET GRANTING SERVER (TGS). A server that provides tickets.
- KEY DISTRIBUTION CENTER (KDC). A server that provides the initial ticket and handles TGS requests; often runs both AS and TGS services.
- TICKET GRANTING TICKET (TGT). The ticket that is granted during the authentication process.
- TICKET. Used to authenticate to the server; contains client identity, sesión key, timestamp, and checksum and is encripted with server's key.
- SESSION KEY. Temporary encryption key.

KERBEROS COMPONENTS







KERBEROS LAB

60 MIN

REFERENCES

1. http://www.roguelynn.com/words/explain-like-im-5-kerberos/