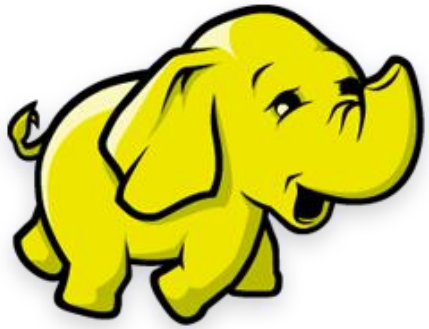
Apache Hadoop Security

COMP9313: Big Data Management

Hadoop and Security



source

- Do we Need Security for Big Data?
- What levels of security do I need?
- Does Hadoop have some security features?

Big Data Security

- Big Data -> Big amount of sensitive information
  - Personal information
  - Intellectual property
  - Trade secrets
  - Financial information
- Huge data breaches
  - Security risks
  - Reputation loss
  - Financial loss

https://www.sciencedirect.com/science/article/pii/S1353485815700097

https://www.sciencedirect.com/science/article/pii/S1353485812700636

Following regulations

Example regulations:

source

- Health Insurance Portability & Accountability Act
- Protection of Personal Identifiable Information from fraud and theft
- Mainly applicable to healthcare and healthcare insurance companies
- Consists of 5 ”Titles”
- Privacy is treated mostly under Title II - *Preventing Health Care Fraud and Abuse*

- General Data Protection Regulation
- EU regulation on data protection and privacy for citizens of the EU and EEA
- Export of data outside EU and EEA
- Gives controls to individuals over their personal data
- Data Controllers -> Must provide technical and organizational measures to implement GDPR
- Hefty fines for violations of GDPR

https://digital.library.unt.edu/explore/collections/CRSR/browse/?q=hipaa&t=fulltext

https://ec.europa.eu/info/law/law-topic/data-protection_en

Honoring Agreements

Big Data SLAs (Service-Level Agreements)

- Companies moving from experimentation to production (mission critical)
- Getting more serious about QoS (Quality of Service)
- Minimum performance and service levels required (e.g. response time and throughput)
- Service-Level Agreements:
  - Commitment between a provider and a user of service

https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1077&context=knoesis

https://www.techrepublic.com/article/how-to-start-using-big-data-slas/

What do I need to Support in regard to Security?

- Authentication
- Authorization
- Encryption
- Monitoring and Auditing

….Enter Hadoop Secure Mode

Authentication

- who is a caller identifying themselves as? and can you verify that they really are this person/service?
- You'd be familiar with Passwords, Face id, fingerprint, secure keys…etc.

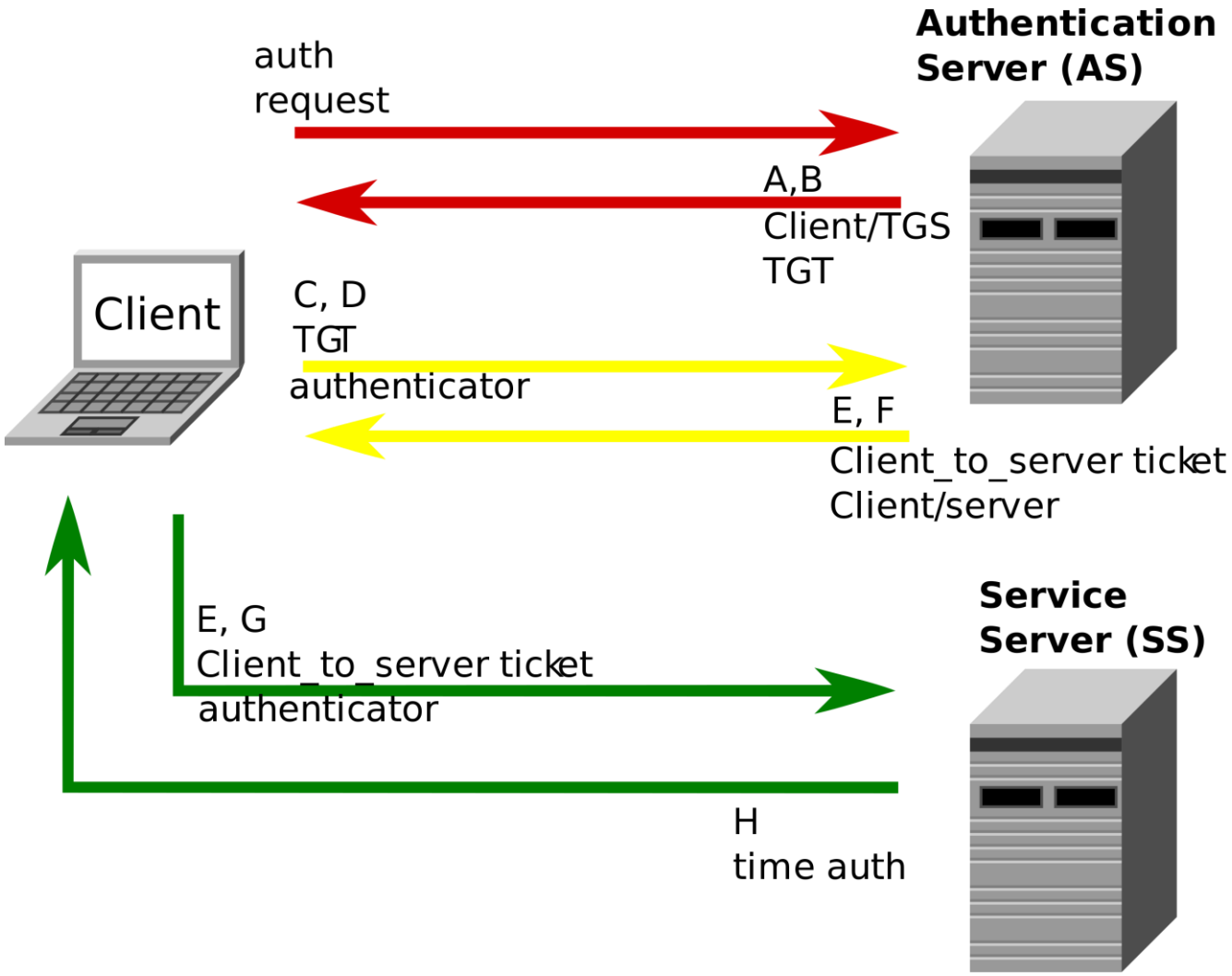
Authentication in Hadoop

- By default… None
- Kerberos was introduces?
- What is Kerberos?

What is Kerberos

- Kerberos is a protocol for authenticating access to distributed services:
  - That callers to a service represent a principal in the system, or
  - That a caller to a service has been granted the right to act on behalf of a principal —a right which the principal can grant for a limited amount of time.

Typical Kerberos flow



Kerberos in Hadoop

1. First thing, you must be authenticated by Kerberos. send an authentication request to Kerberos Authentication Server.

2. On a successful authentication, the Authentication Server will respond back with a Ticket Granting Ticket TGT.

3. So, now you have your TGT that means, you have got your authentication, and you are ready to execute a Hadoop command.

4. Let's say you run following command.
hadoop fs --ls /
So, you are using Hadoop command.

Kerberos in Hadoop (Cont'd)

5. Now, the Hadoop client will use your TGT and reach out to Ticket Granting Service TGS. The client approaches TGS to ask for a service ticket for the Name Node service.

6. The TGS will grant you a service ticket, and the client will cache the service ticket.

7. Now, you have a ticket to communicate with the Name Node. So, the Hadoop RPC will use the service ticket to reach out to Name Node.

8. They will again exchange the tickets. Your Ticket proves your identity and Name node's Ticket determines the identification of the Name Node. Both are sure that they are talking to an authenticated entity. We call this a mutual authentication.

Authorization

- Does an (authenticated) user/service have the permissions to perform the desired request?
- You might be familiar with Access Control, ACL in firewalls, permissions in your phone…etc.

Authorization in Hadoop

- This isn't handled by Keberos: this is Hadoop-side, and is generally done in various ways across systems.
- HDFS has file and directory permissions, with the user+group model now extended to ACLs.
- YARN allows job queues to be restricted to different users and groups, so restricting the memory & CPU limits of those users. When cluster node labels are used to differentiate parts of the cluster (e.g. servers with more RAM, GPUs or other features), then the queues can be used to restrict access to specific sets of nodes.

Encryption

- Can data be intercepted on disk or over the wire?
- Things to think about?
  - Cost
  - Performance
  - Scale
  - Protect what need protection

Encryption in Hadoop

- HDFS now supports at rest encryption; the data is encrypted while stored on disk.
- Before rushing to encrypt all the data, consider that it isn't a magic solution to security: the authentication and authorisation comes first.
- Encryption adds a new problem, secure key management, as well as the inevitable performance overhead. It also complicates some aspects of HDFS use.

Encryption in Hadoop

- Setting **hadoop.rpc.protection** to **privacy** in **core-site.xml** activates data encryption on Rest.
- You need to set **dfs.encrypt.data.transfer** to **true** in the **hdfs-site.xml** in order to activate data encryption for data transfer protocol of DataNode
- Optionally, you may set **dfs.encrypt.data.transfer.algorithm** to either **3DES** or **RC4** to choose the specific encryption algorithm. If unspecified, then the configured JCE default on the system is used, which is usually 3DES.

Auditing and Monitoring

- Authenticated and Authorized users should not just be able to perform actions or read and write data this should all be logged in Audit Logs.
- Administrators should be able to tell what is happening in the system and who is doing what.

Auditing and Monitoring in Hadoop

- Jobs on the NameNodes and JobTrackers should be logged
- Authorization Failure should be logged
- Authentication Failures should be logged

Useful Security tools for Hadoop

- Apache Ranger
- Apache Sentry
- Apache Knox

Questions?

Notes

- Thursday lecture time we'll have a hands-on activity
  - Bring Laptop
  - We'll run a step by step activity
  - We can't expect for everyone to complete in time that is why we have Labs