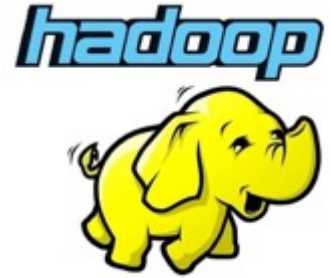
Hadoop Security Topics

by Timothy Spann (@PaasDev)

http://sparkdeveloper.com/

Hadoop Security Topics



- What is Hadoop?
- Overview
- Versioning
- Encryption
- Default Passwords
- Admin HTTP Sites
- Ports
- Threat Vectors
- Google Hacks
- Solutions (Knox, Ranger)

What is Hadoop?

Hadoop consists of an alphabet list of associate Apache projects (Pig, Hive, HBase, Spark, Flume, Zookeeper, Sqoop, Oozie, Storm and others). Hadoop is implemented in Java with some side projects in Scala.

Hadoop consists of core libraries, a distributed file system (HDFS), YARN for resource management and application scheduling, and Map Reduce for data processing. It is an open sourced implementation of some of Google's technologies. It is widely used, cloud hosted and available in many versions and distributions from many vendors.

Hadoop Security

- Hadoop data lakes are everywhere
- Unencrypted raw files of many sources are present
- Often default downloads of open source is used
- Large number of data nodes can lead to misconfigured security and firewalls
- Lazy admins often leave defaults
- Multiple sources of ingest leads to multiple attack vectors
- MySQL is often used

Hadoop Versioning

In pre-2.6 versions of Hadoop, DataNodes were secured by using privileged ports and running services as Root.

There are many versions of Hadoop and numerous distributions, each containing many subproject, libraries and tools. It is very easy to fall behind on versions, patches and systems.

Many are running early versions with ancient Java 6 JVM to run them.

Hadoop Misconfiguration

Not only is misconfiguration easy, running old versions of part or all of the platform is possible if not using a properly installed distribution like HDP or CDH. Options like **hadoop.rpc.protection** in **core-site.xml** have to be manually set and default to off. There are many steps that need to be followed for security to be properly setup, testing is critical.

Hadoop Data Ingest

Often ports are left open, no passwords used and large files blindly accepted and processed. Outside web sites and APIs are called and those could be compromised.

By default security is off, it needs to be enabled, configured and tested. Patches must be maintained as versions often change. Each module has it's own versioning, often below 1.0. Pig is 0.15.0, Hive 1.21, HBase 1.1.1, Sqoop 1.4.6, Flume 1.5.2, Kafka 0.8.2, Knox 0.6.0, Ranger 0.5.0. There are dozens of other tools, frameworks and projects in the Hadoop ecosystem that are often used. Your best choice is to use Apache NiFi with SSL.

https://cwiki.apache.org/confluence/display/SQOOP/Security+Guide+On+Sqoop+2

Hadoop Data in Motion

3DES and RC4 are often used, but have to be configured and are for Block Data Transfer. RPCs have another security configuration and HTTP / REST calls use SSL that must be enabled for a litany of services.

Hadoop Data At Rest

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/TransparentEncryption.html

In 2.7.1 the newest release this year, it can be enabled.

HDFS

Hadoop file system security relies on proper UNIX style permissions and underlying operating systems. Access is available via Web API and command line tool. Requires careful ACL setup and is on top of Linux, usually RHEL.

Accumulo

HBase Database Security

Developed by NSA

Cell level authorization

Hive

Hive Database Security

GRANT/REVOKE SQL permissions

This requires the same attention that an enterprise RDBMS does to ensure that the access control to data is tightly controlled. Hive can also be accessing JSON, raw Text files and unstructured types of data that may have exposed credit cards, personal data or anything else that may be on the web. This data could contain links to malware or could be illegal.

Hadoop Default Passwords

Apache Ranger: admin/admin
HDP Security Admin: admin/admin
SSH / Root: root/hadoop
Hue: hue/1111
Atlas: admin/admin
Falcon: falcon

Most services have easy to guess defaults, no password or no login at all. Often default installed services may be left open and provide a backdoor to HDFS allowing unfettered access to all files and data.

http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/#section_6

Hadoop Admin HTTP Sites

Ambari Console:
http://hadoopnode:8080

Ambari is the main configuration and administration console for all of Hadoop.

Apache Ranger:
http://hadoopnode:6080

Each service and node has it’s own administration site that may or may not be secured.

Hadoop Default Ports

50070 : Name Node Web UI
50470 : Name Node HTTPS Web UI
8020, 8022, 9000 : Name Node via HDFS
50075 : Data Node(s) WebUI
50475 : Data Node(s) HTTPS Web UI
50090 : Secondary Name Node
60000 : HBase Master
8080 : HBase REST
9090 : Thrift Server
50111 : WebHCat
8005 : Sqoop2
2181: Zookeeper
9010: Zookeeper JMX

More Ports

50020, 50010, 50030, 8021, 50060, 51111
9083, 10000, 60010, 60020, 60030
2888, 3888, 8660, 8661, 8662, 8663, 8660, 8651
3306, 80, 8085, 1004, 1006, 8485, 8480, 2049, 4242
14000, 14001, 8021, 9290, 50060, 8032, 8030, 8031
8033, 8088, 8040, 8042, 8041, 10020, 13562, 19888
9090, 9095, 9083, 16000, 12000, 12001, 3181, 4181,
8019, 9010, 8888, 11000, 11001, 7077, 7078, 18080, 18081, 50100

Threat Vectors

- **Web Application Security**
- **Network Security**
- **Social Engineering**
- **Patches**
- **Password Exposure**
- **Weak Keys**

Threat Vectors

PasteBin Leaks

Patches
Not patching OS, SFTP, FTP, Firewalls, SSH, SSL, Hadoop, Spark, Hive, Pig, Zookeeper, Java, …

Java 6 JVM usage (http://www.darkreading.com/vulnerabilities-and-threats/hackers-target-java-6-with-security-exploits/d/d-id/1111293?) Even some that are Java 7 still have risks as both versions are EOL.

SSH “edge nodes” to gain access to the cluster (if not using Knox).

Google Hacks

filetype:log intext:org.apache.hadoop.hdfs

inurl:dfshealth.jsp

inurl:dfshealth.html

Google Hacks

inurl:dfshealth.html

Open Node on the Internet without Passwords

inurl:”50070/logs/"

Google Hacks

inurl:”50070/logs/"

inurl:”8088/cluster"

Shodan Search

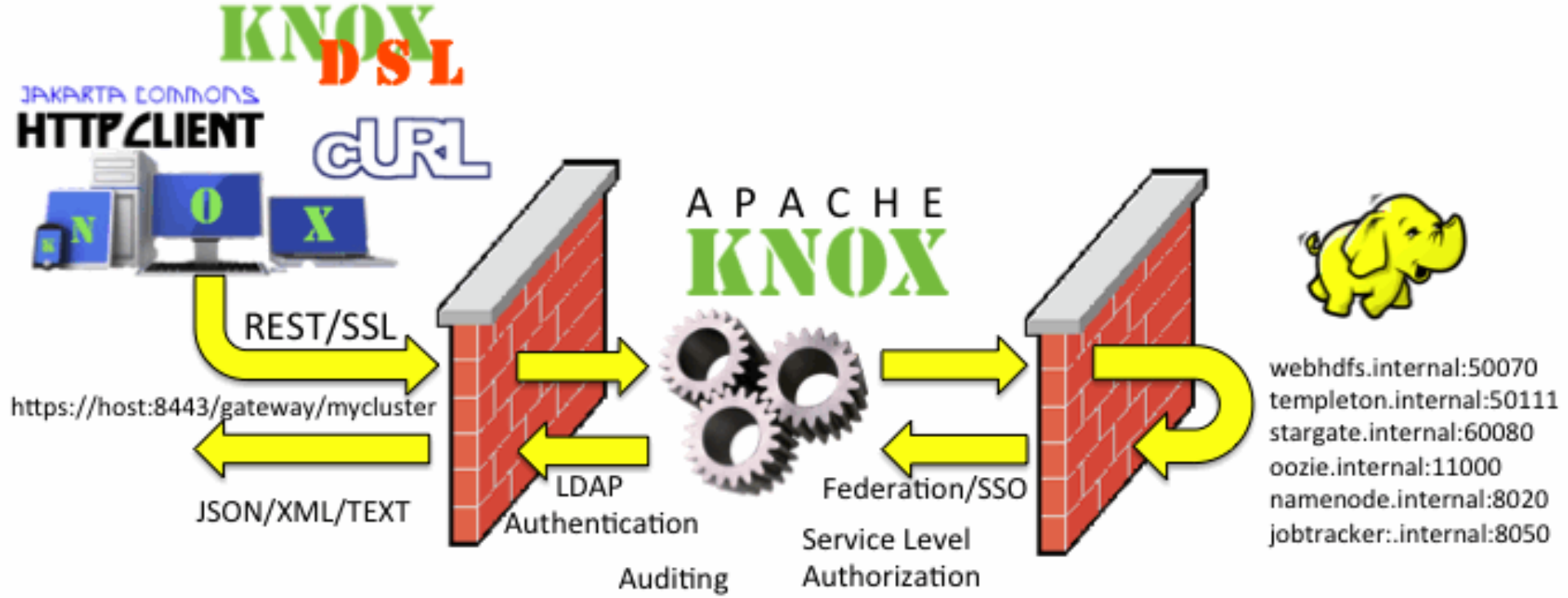
port:16010 product:”HBase"

Security Solutions

Apache Knox is a gateway to access a Hadoop cluster for data access and job control. Knox has been available for over 3 years.

- REST/HTTP services encapsulating Kerberos
- Enforces REST API security
- Supports LDAP, ActiveDirectory, SSO, OAuth, OpenID, MFA and SAML
- Perimeter security
- Stateless reverse proxy framework

- Knox uses Shiro authentication provider by default against LDAP.
- Knox by default uses ACL based.
- Auditing to Log4j
- Supports WebHDFS, HCatalog, HBase, Oozie, Hive/JDBC, YARN, Storm


KNOX
DSL
JAKARTA COMMONS
HTTPCLIENT
cURL
N O X
REST/SSL
https://host:8443/gateway/mycluster
JSON/XML/TEXT
APACHE
KNOX
LDAP
Authentication
Auditing
Federation/SSO
Service Level
Authorization
webhdfs.internal:50070
templeton.internal:50111
stargate.internal:60080
oozie.internal:11000
namenode.internal:8020
jobtracker:.internal:8050

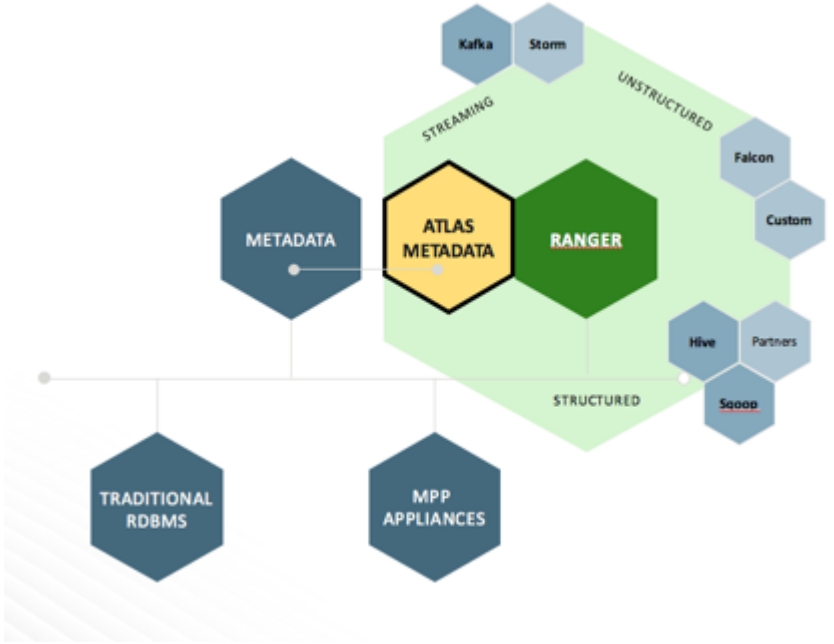
Apache Ranger

Apache Ranger is designed for central security policy administration for a Hadoop cluster. It provides authorization, accounting and data protection.

Policies can be setup universally or per user or group, using a GRANT syntax familiar to all RDBMS DBAs.

Data Governance

- Auditing
- Compliance with HIPAA, PCI and others around PII (personally identifiable data)
- Authentication
- Authorization
- Data Protection



Data Governance
http://hortonworks.com/apache/atlas/

Hadoop in Secure Mode

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SecureMode.html

The process is documented and must be followed carefully and include the steps for securing underlying Linux servers.

Hadoop Security Links

http://www.datanami.com/2014/11/06/big-data-breach-security-concerns-still-shadow-hadoop/
http://www.datanami.com/2014/06/24/hadoop-security-still-lot-work/
http://hortonworks.com/hdp/security/
http://data-informed.com/how-to-secure-data-in-hadoop/
http://www.infoworld.com/article/2849679/application-development/hadoop-security-apache-ranger.html
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html
http://www.esecurityplanet.com/network-security/hadoop-security-still-evolving.html
http://www.securityweek.com/bigger-data-smaller-problems-managing-security-permissions-data-subsets-hadoop
http://www.securityweek.com/big-data-smaller-problems-configuring-kerberos-authentication-hadoop

Hadoop Security Links

https://accumulo.apache.org/1.4/user_manual/Security.html
http://www.infoq.com/articles/HadoopSecurityModel
https://github.com/intel-hadoop/project-rhino/
http://spark.apache.org/docs/latest/security.html
http://www.slideshare.net/DonaldMiner/accumulo-oct2013bofpresentation
http://hbase.apache.org/0.94/book/security.html
http://hortonworks.com/hadoop-tutorial/securing-data-lake-auditing-user-access-using-hdp-security/
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SecureMode.html

Hadoop Security Links

http://www.slideshare.net/Hadoop_Summit/t-325210-anaurothv8mac
http://www.slideshare.net/Hadoop_Summit/th-130p211minder
http://www.slideshare.net/oom65/hadoop-security-architecture?next_slideshow=1
https://blog.samanage.com/insights/big-data-is-your-it-service-desk-informed-on-hadoop-security/
http://hortonworks.com/blog/hadoop-security-enterprise/
http://blogs.gartner.com/merv-adrian/2014/01/21/security-for-hadoop-dont-look-now/

Hadoop Security Links - Knox

http://hortonworks.com/hadoop/knox-gateway/#tutorials
http://hortonworks.com/hadoop-tutorial/manage-security-policy-hive-hbase-knox-ranger/
https://cwiki.apache.org/confluence/display/KNOX/Examples+Hive
http://hortonworks.com/hadoop-tutorial/securing-hadoop-infrastructure-apache-knox/
http://hortonworks.com/hadoop-tutorial/secure-jdbc-odbc-clients-access-hiveserver2-using-apache-knox/

Hadoop Links

http://ambari.apache.org/
https://en.wikipedia.org/wiki/Apache_Hadoop
https://ambari.apache.org/1.2.3/installing-hadoop-using-ambari/content/reference_chap2_1.html
http://hortonworks.com/blog/hadoop-summit-curated-content-apache-hadoop-security/