Apache Spark & Apache Zeppelin: Enterprise Security for production deployments

Vinay Shukla
Director, Product Management
Nov 15, 2016
Twitter: @neomythos

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Thank You

HORTONWORKS®
POWERING THE FUTURE OF DATA™

whoami

Recovering Programmer, Product Management

- Product Management
- Spark for 2.5 + years, Hadoop for 3+ years
- Blog at www.vinayshukla.com
- Twitter: @neomythos
- Addicted to Yoga, Hiking, & Coffee
- Smallest contributor to Apache Zeppelin

HORTONWORKS

What are the security requirements?

- Spark user should be authenticated
- Integrate with corporate LDAP/AD
- Allow only authorized users access
- Audit all access
- Protect data both in motion & at rest
- Easily manage all security
- Make security easy to manage
- …

HORTONWORKS

Security: Rings of Defense

Perimeter Level Security

- Network Security (i.e. Firewalls)

Authentication

- Kerberos
- Knox (Other Gateways)

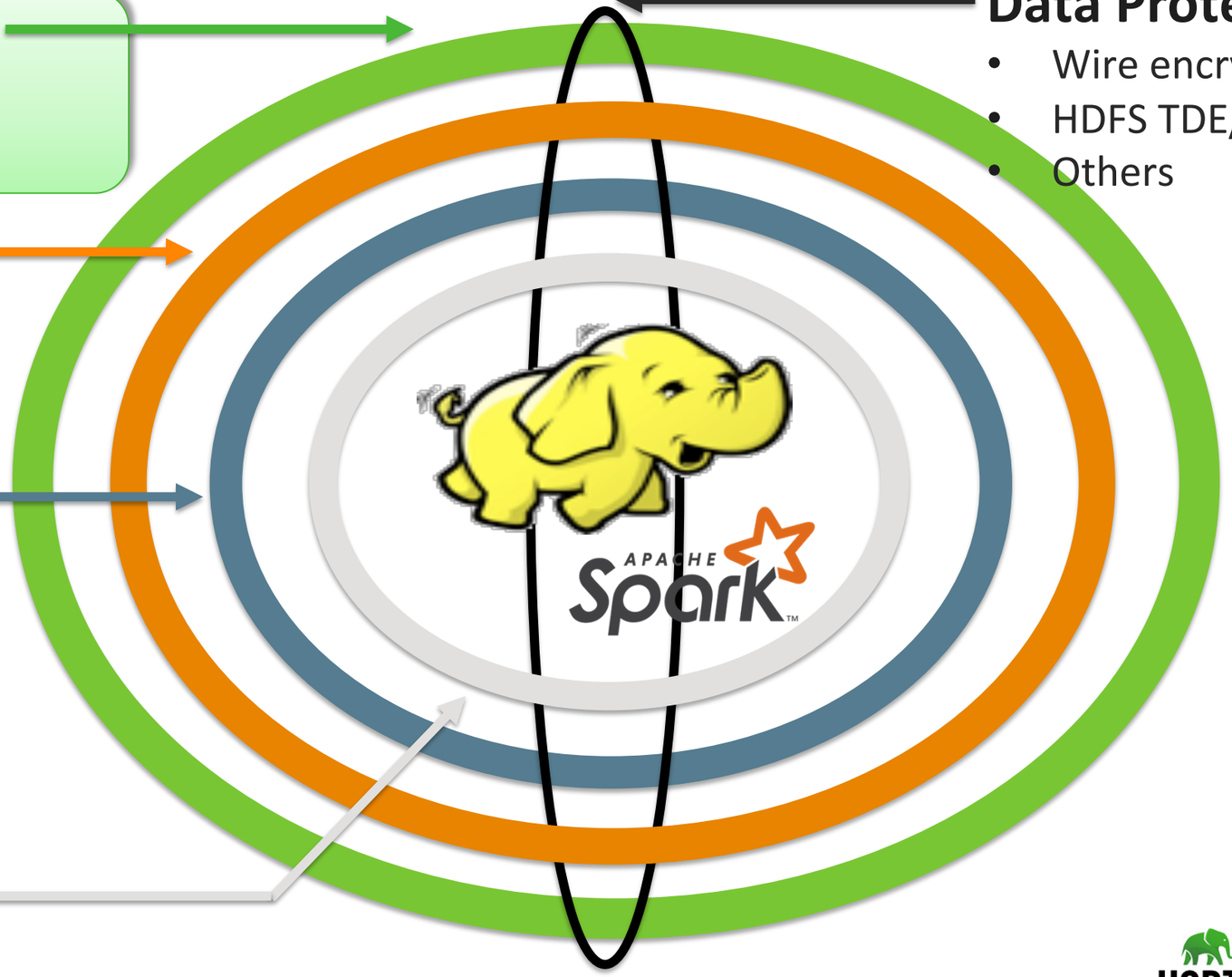
Authorization

- Apache Ranger/Sentry
- HDFS Permissions
- HDFS ACLs
- YARN ACL

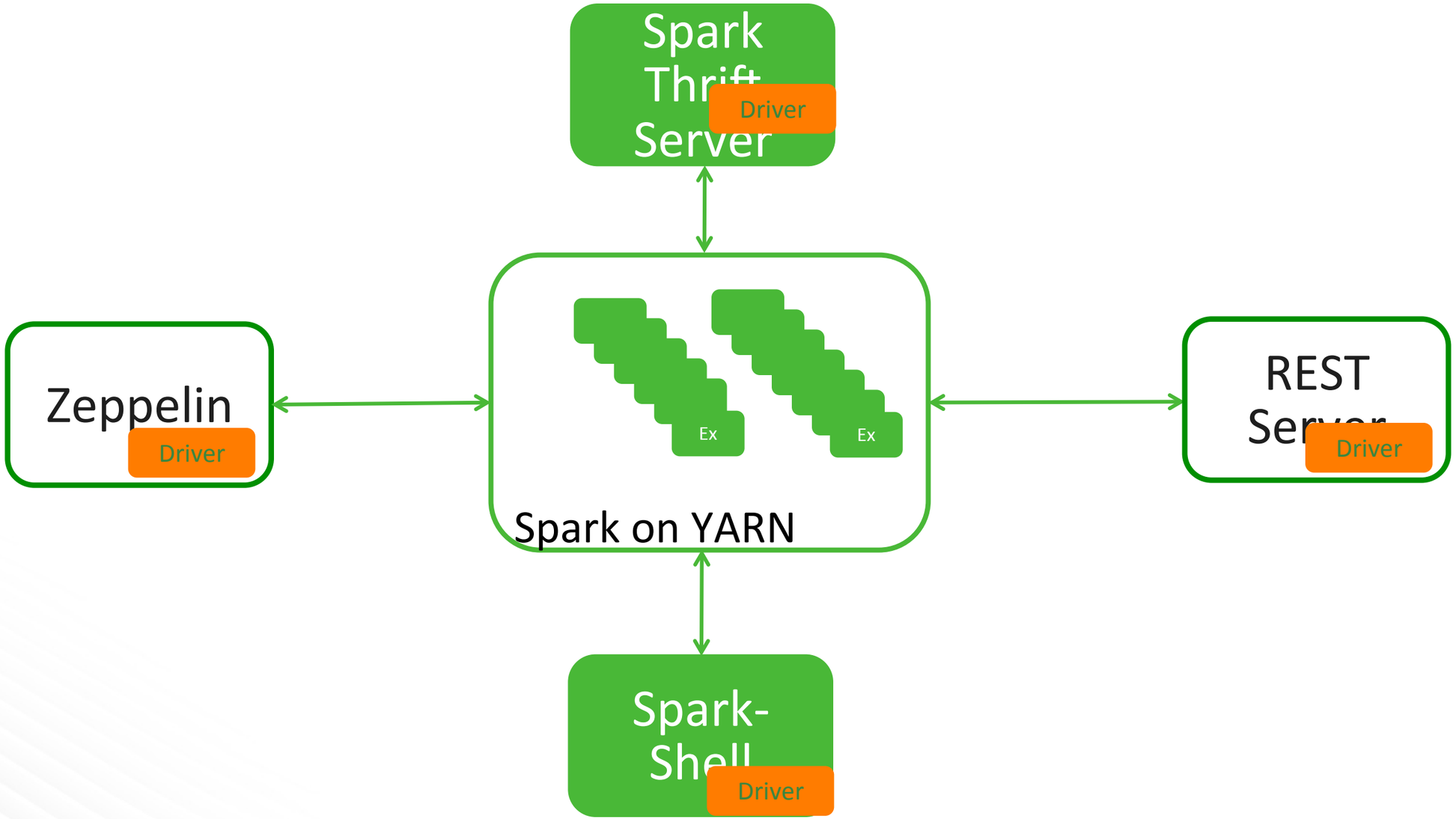
OS Security

Data Protection

- Wire encryption
- HDFS TDE/DARE
- Others



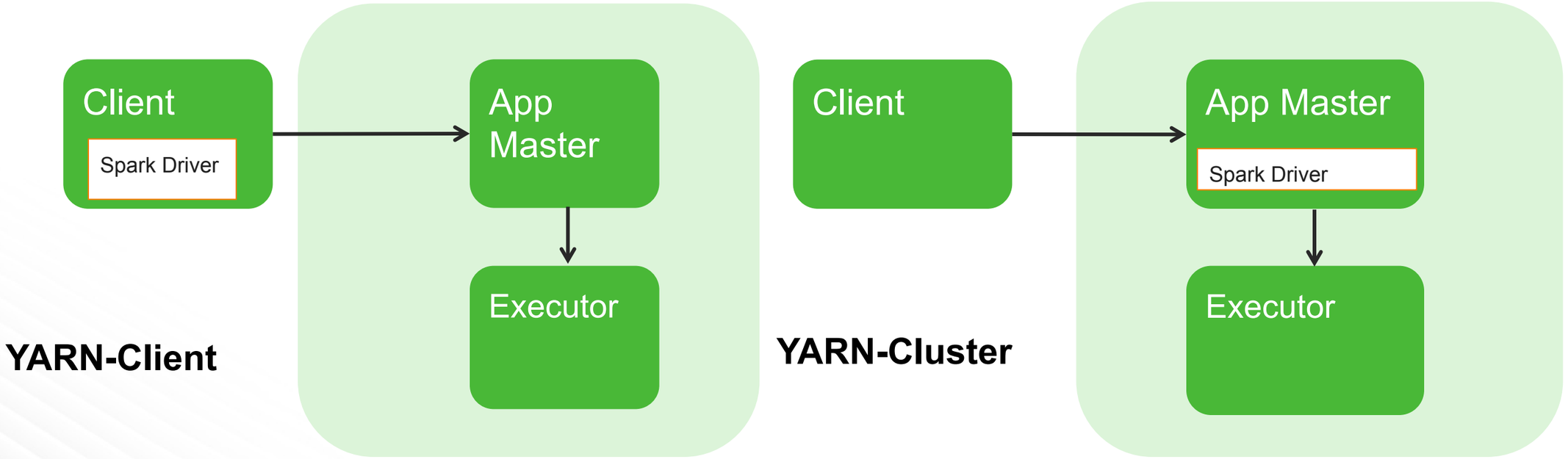
Interacting with Spark



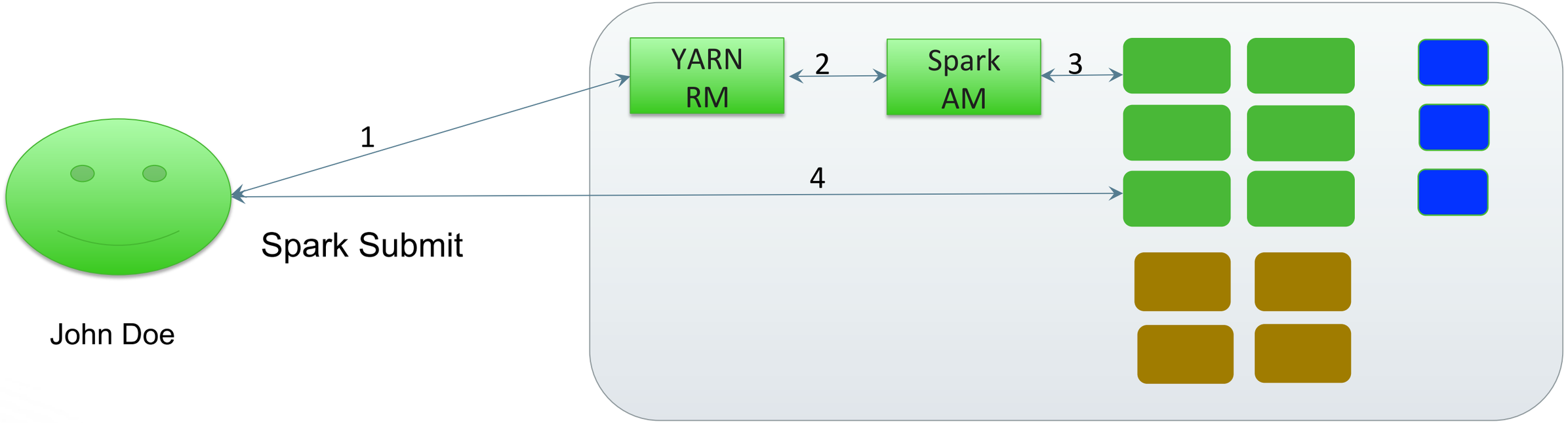
HORTONWORKS

Context: Spark Deployment Modes

- Spark on YARN
  - Spark driver (SparkContext) in YARN AM(yarn-cluster)
  - Spark driver (SparkContext) in local (yarn-client):
    - Spark Shell & Spark Thrift Server runs in yarn-client only



Spark on YARN



HORTONWORKS

Spark – Security – Four Pillars

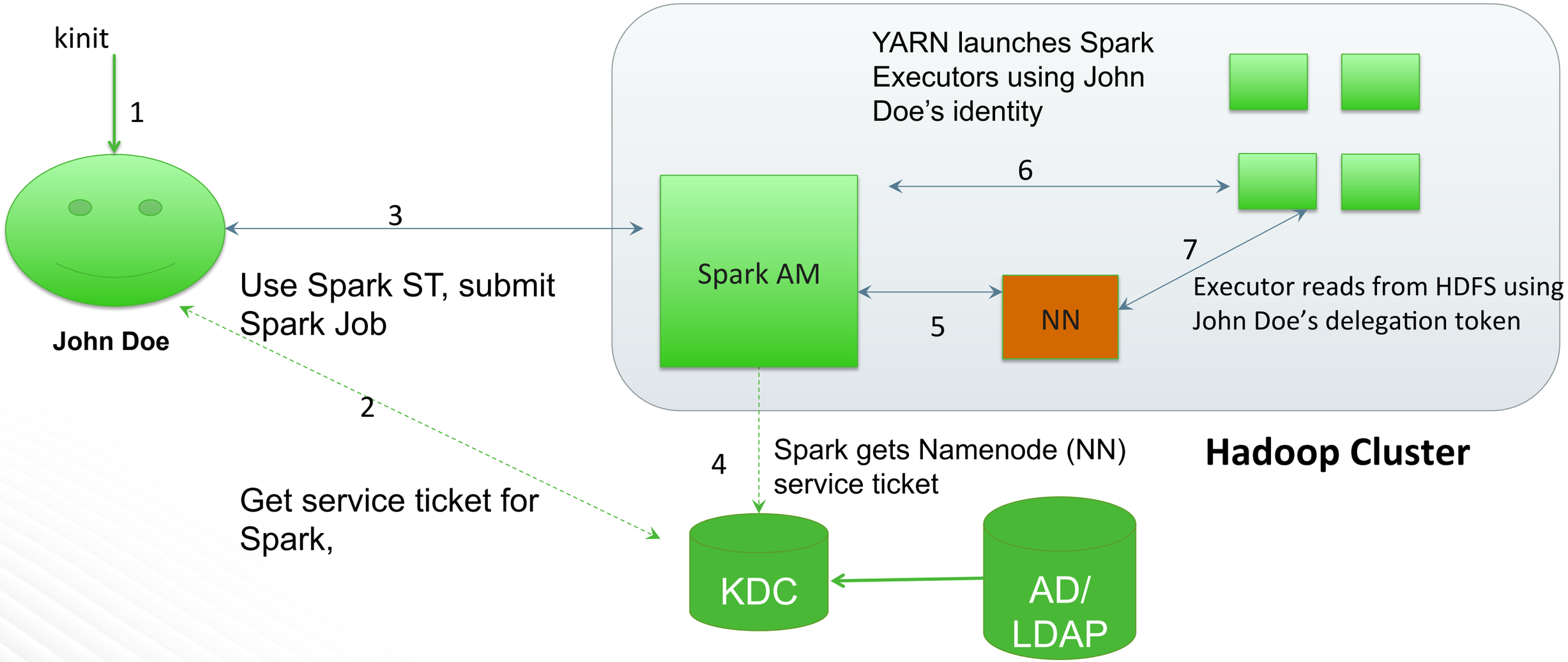
- Authentication
- Authorization
- Audit
- Encryption

Ensure network is secure

Spark leverages Kerberos on YARN

HORTONWORKS

Authenticate users with AD/LDAP



HORTONWORKS

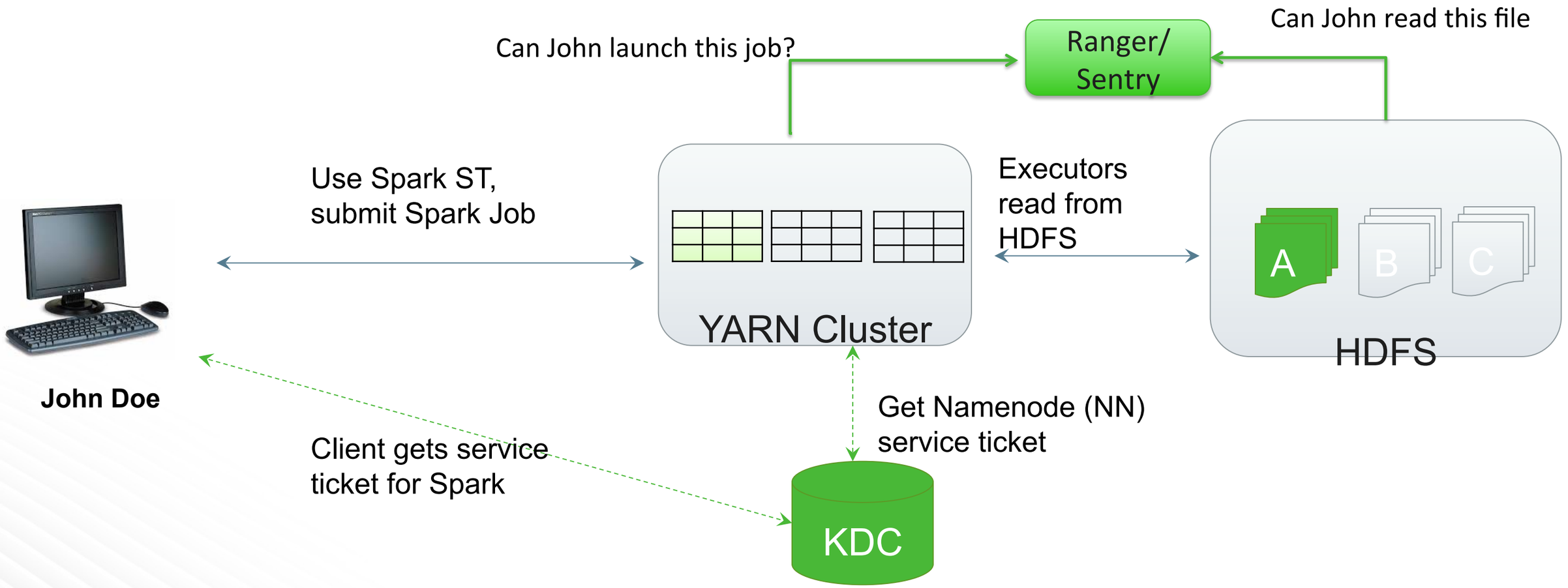
Spark – Kerberos - Example

```
kinit -kt /etc/security/keytabs/johndoe.keytab
johndoe@EXAMPLE.COM
```

```
./bin/spark-submit --class org.apache.spark.examples.SparkPi
--master yarn-cluster  --num-executors 3 --driver-memory 512m
--executor-memory 512m   --executor-cores 1  lib/spark-
examples*.jar 10
```

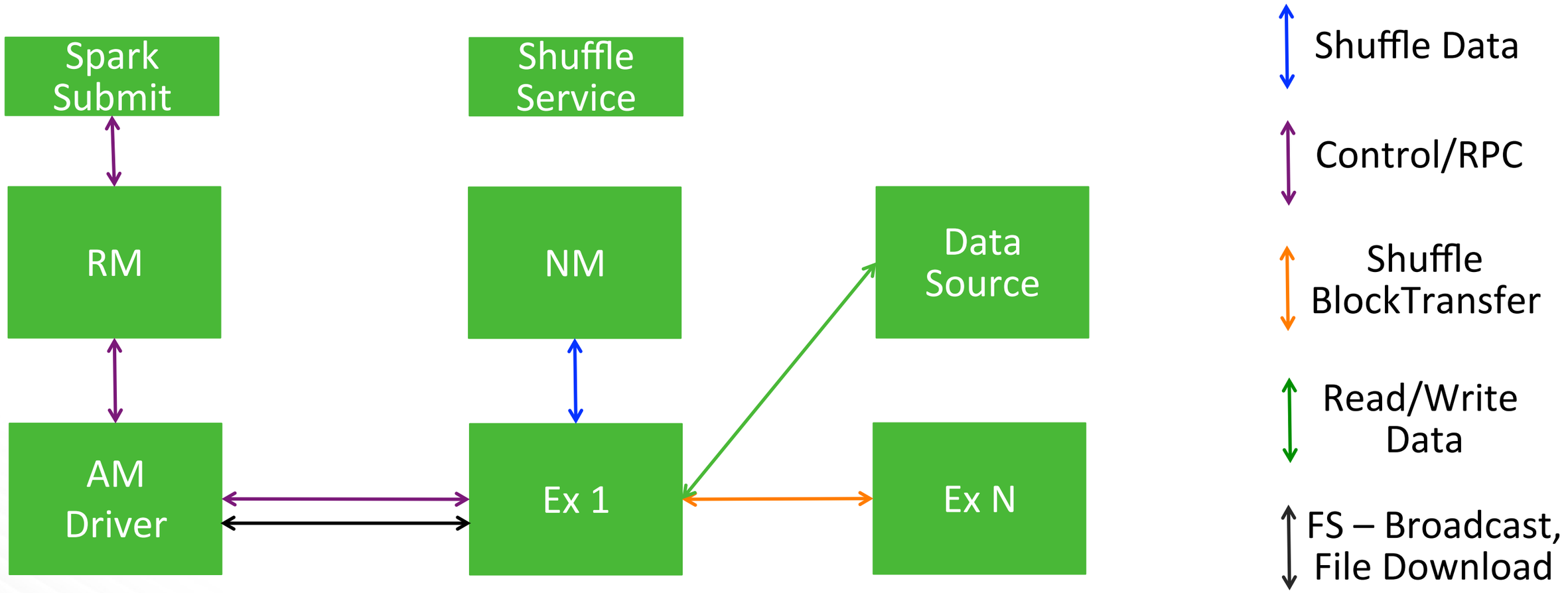

HORTONWORKS

Allow only authorized users access to Spark jobs



HORTONWORKS

Secure data in motion: Wire Encryption with Spark



HORTONWORKS

Spark Communication Encryption Settings

Shuffle Data

NM > Ex leverages YARN based SSL

Control/RPC

spark.authenticate = true. Leverage YARN to distribute keys

Shuffle BlockTransfer

spark.authenticate.enableSaslEncryption= true

Read/Write Data

Depends on Data Source, For HDFS RPC (RC4 | 3DES) or SSL for WebHDFS

FS – Broadcast, File Download

spark.ssl.enabled = true

HORTONWORKS

Sharp Edges with Spark Security

- SparkSQL – Only coarse grain access control today
- Client -> Spark Thrift Server > Spark Executors – No identity propagation on 2nd hop
  - Lowers security, forces STS to run as Hive user to read all data
  - Use SparkSQL via shell or programmatic API
  - https://issues.apache.org/jira/browse/SPARK-5159
- Spark Stream + Kafka + Kerberos
  - No SSL support yet
- Spark Shuffle > Only SASL, no SSL support
- Spark Shuffle > No encryption for spill to disk or intermediate data

HORTONWORKS

SparkSQL: Fine grained security

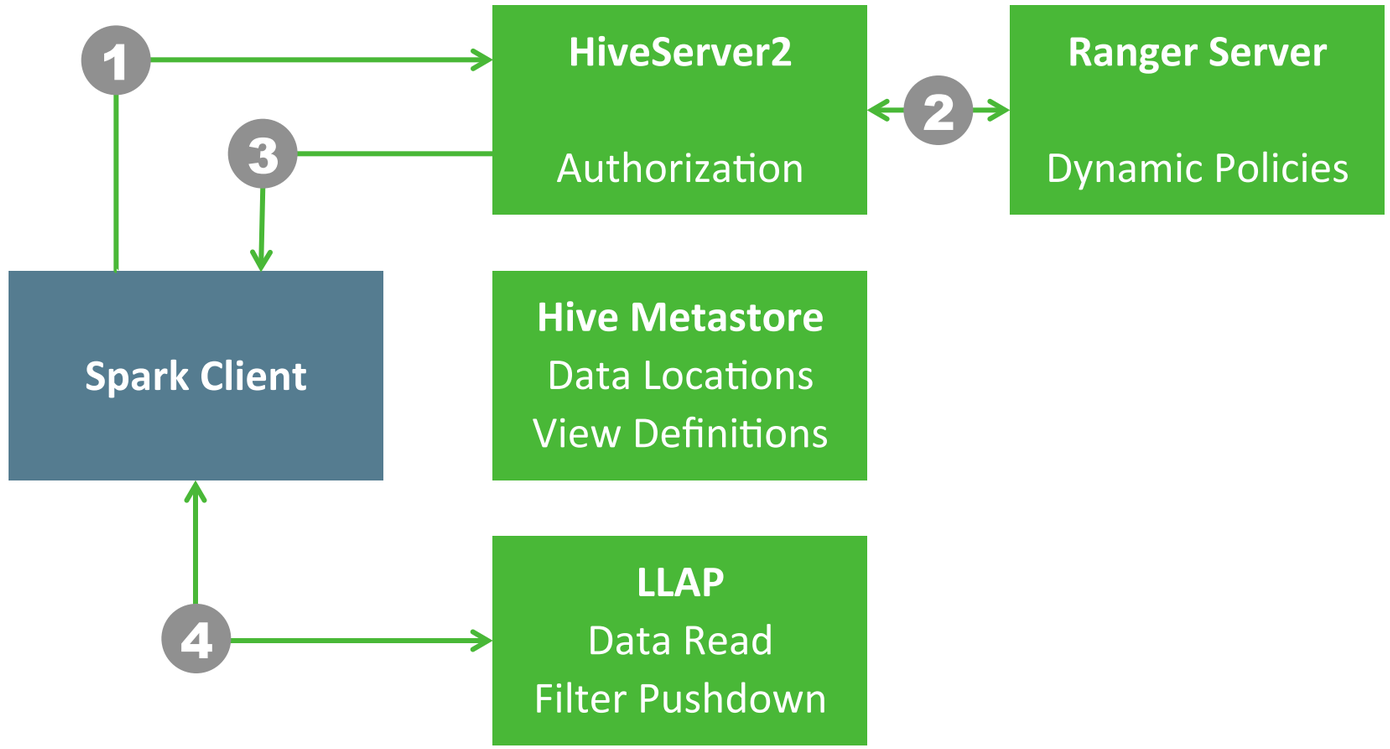
HORTONWORKS

Key Features: Spark Column Security with LLAP

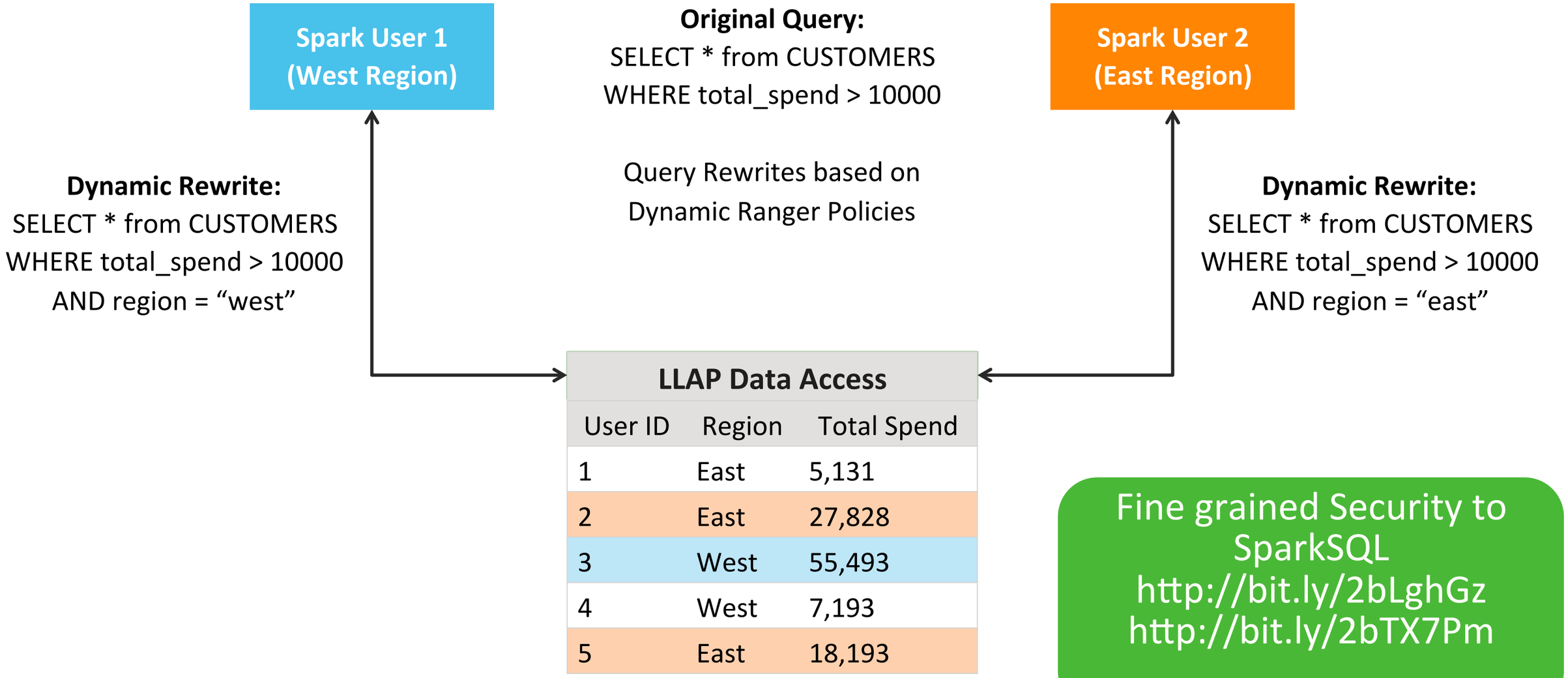
- Fine-Grained Column Level Access Control for SparkSQL.
- Fully dynamic policies per user. Doesn't require views.
- Use Standard Ranger policies and tools to control access and masking policies.

Flow:

1. SparkSQL gets data locations known as "splits" from HiveServer and plans query.
2. HiveServer2 authorizes access using Ranger. Per-user policies like row filtering are applied.
3. Spark gets a modified query plan based on dynamic security policy.
4. Spark reads data from LLAP. Filtering / masking guaranteed by LLAP server.



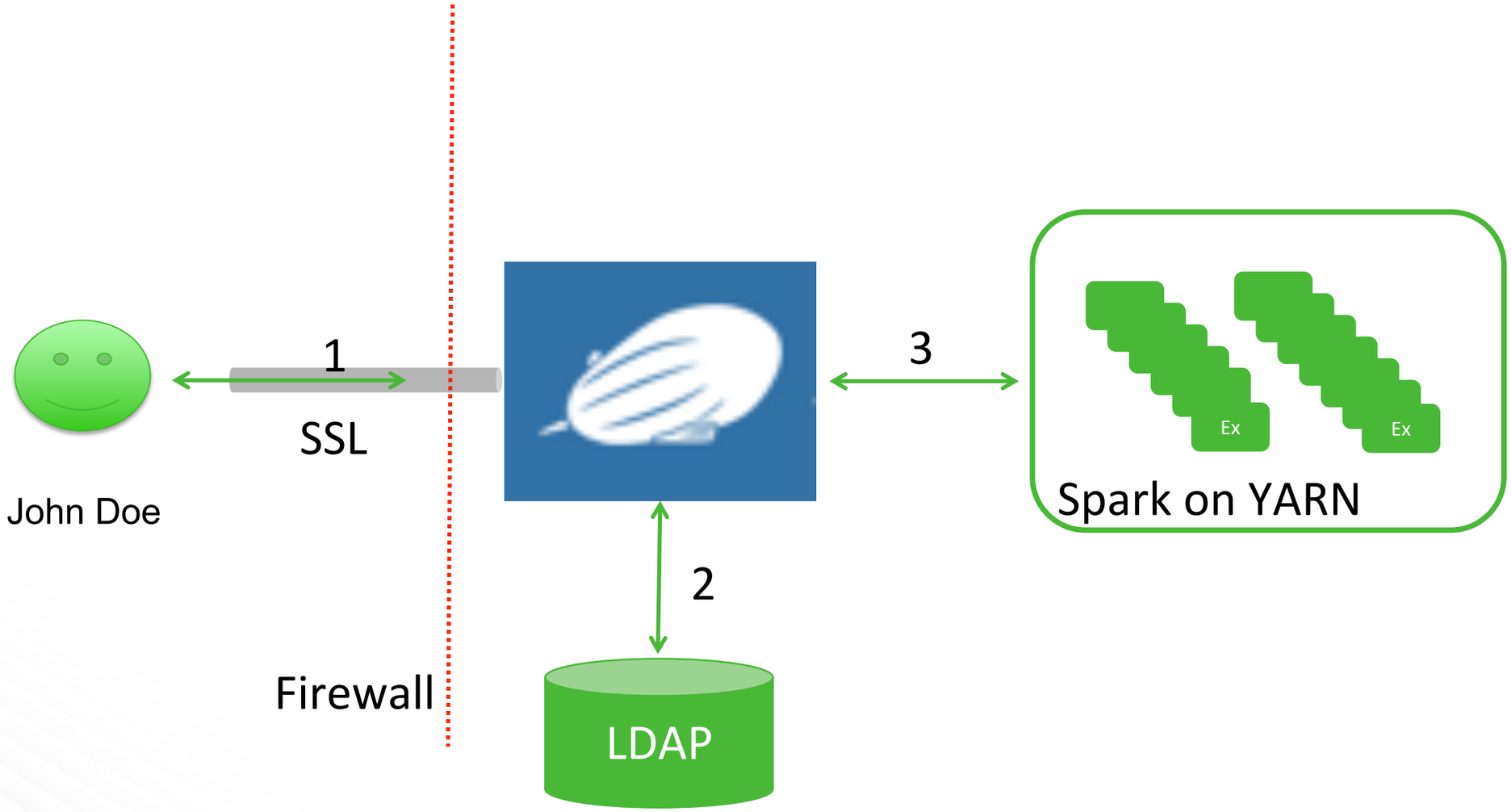
Example: Per-User Row Filtering by Region in SparkSQL



Apache Zeppelin Security

HORTONWORKS

Apache Zeppelin: Authentication + SSL



HORTONWORKS

Security in Apache Zeppelin?

Zeppelin leverages Apache Shiro for authentication/ authorization

HORTONWORKS

Example Shiro.ini

```
# =======================
# Shiro INI configuration
# =======================

[main]
## LDAP/AD configuration

[users]
# The 'users' section is for simple deployments
# when you only need a small number of statically-defined
# set of User accounts.

[urls]
# The 'urls' section is used for url-based security
#
```

Edit with Ambari or your favorite text editor

HORTONWORKS

Apache Zeppelin: AD Authentication

- Configure Zeppelin to use AD

```
[main]
activeDirectoryRealm = org.apache.zeppelin.server.ActiveDirectoryGroupRealm
activeDirectoryRealm.systemUsername = XXXXX
activeDirectoryRealm.systemPassword = XXXXXXXXXXXXXXXXX
activeDirectoryRealm.searchBase = DC=hdpqa,DC=Example,DC=com
activeDirectoryRealm.url = ldap://hdpqa.example.com:389
activeDirectoryRealm.principalSuffix = @hdpqa.example.com
activeDirectoryRealm.groupRolesMap =
"CN=hdpdv_admin,DC=hdpqa,DC=example,DC=com":"admin"
activeDirectoryRealm.authorizationCachingEnabled = true
sessionManager = org.apache.shiro.web.session.mgt.DefaultWebSessionManager
cacheManager = org.apache.shiro.cache.MemoryConstrainedCacheManager
securityManager.cacheManager = $cacheManager
securityManager.sessionManager = $sessionManager
securityManager.sessionManager.globalSessionTimeout = 86400000
shiro.loginUrl = /api/login
```


HORTONWORKS

Apache Zeppelin: LDAP Authentication

- Configure Zeppelin to use LDAP

```
[main]
ldapRealm = org.apache.zeppelin.server.LdapGroupRealm
ldapRealm = org.apache.shiro.realm.ldap.JndiLdapRealm
ldapRealm.contextFactory.environment[ldap.searchBase] =
DC=hdpqa,DC=example,DC=com
ldapRealm.userDnTemplate = uid={0},OU=Accounts,DC=hdpqa,DC=example,DC=com
ldapRealm.contextFactory.url = ldaps://hdpqa.example.com:636
ldapRealm.contextFactory.authenticationMechanism = SIMPLE
sessionManager = org.apache.shiro.web.session.mgt.DefaultWebSessionManager
securityManager.sessionManager = $sessionManager
# 86,400,000 milliseconds = 24 hour
securityManager.sessionManager.globalSessionTimeout = 86400000
shiro.loginUrl = /api/login
```


HORTONWORKS

Don’t want passwords in clear in shiro.ini?

- Create an entry for AD credential
  - Zeppelin leverages Hadoop Credential API
  - `hadoop credential create`

```
activeDirectoryRealm.systemPassword -provider jceks:///etc/zeppelin/conf/
credentials.jceks
Enter password:
Enter password again:
activeDirectoryRealm.systemPassword has been successfully created.
org.apache.hadoop.security.alias.JavaKeyStoreProvider has been updated.
```

- Make credentials.jceks only Zeppelin user readable
  - chmod 400 with only Zeppelin process r/w access, no other user allowed access to Credentials
- Edit shiro.in
  - activeDirectoryRealm.systemPassword -provider jceks://etc/zeppelin/conf/credentials.jceks

HORTONWORKS

Want to connect to LDAP over SSL?

- Change protocol to ldaps in shiro.ini

```
ldapRealm.contextFactory.url = ldaps://hdpqa.example.com:636
```

- If LDAP is using self signed certificate, import the certificate into truststore of JVM running Zeppelin

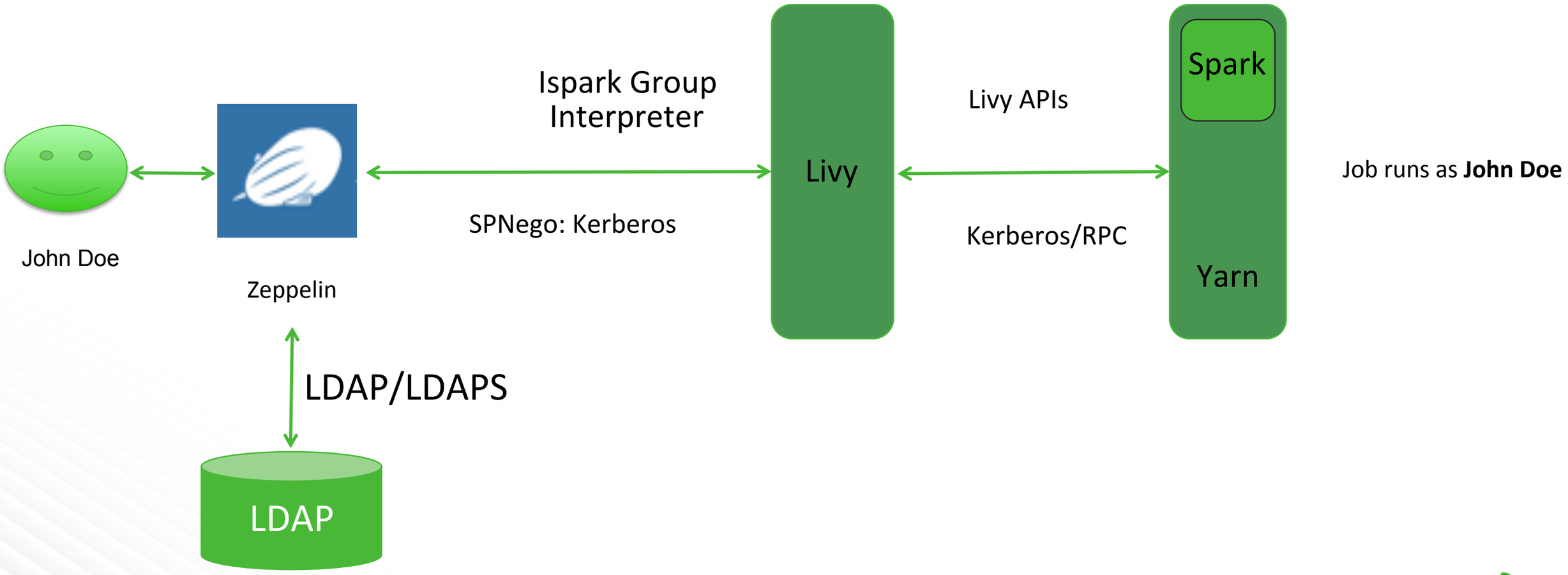
```
echo -n | openssl s_client –connect ldap.example.com:389 | \
   sed -ne '/-BEGIN CERTIFICATE-/,/-END CERTIFICATE-/p' > /tmp/
examplecert.crt

keytool –import -keystore $JAVA_HOME/jre/lib/security/cacerts \

 -storepass changeit -noprompt -alias mycert -file /tmp/examplecert.crt
```


HORTONWORKS

Zeppelin + Livy E2E Security



HORTONWORKS

Apache Zeppelin: Authorization

Authorization in Zeppelin

- Note level authorization
  - Grant Permissions (Owner, Reader, Writer) to users/groups on Notes
  - LDAP Group integration
- Zeppelin UI Authorization
  - Allow only admins to configure interpreter
  - Configured in shiro.ini

```
[urls]
/api/interpreter/** = authc, roles[admin]

/api/configurations/** = authc, roles[admin]

/api/credential/** = authc, roles[admin]
```

Authorization at Data Level

- For Spark with Zeppelin > Livy > Spark
  - Identity Propagation Jobs run as End-User
- For Hive with Zeppelin > JDBC interpreter
- Shell Interpreter
  - Runs as end-user

HORTONWORKS

Map admin role to AD Group

- Allows mapped AD group access to Configure Interpreters

```
[main]
activeDirectoryRealm = org.apache.zeppelin.server.ActiveDirectoryGroupRealm
activeDirectoryRealm.systemUsername = XXXXX
activeDirectoryRealm.systemPassword = XXXXXXXXXXXXXXXXX
activeDirectoryRealm.searchBase = DC=hdpqa,DC=Example,DC=com
activeDirectoryRealm.url = ldap://hdpqa.example.com:389
activeDirectoryRealm.principalSuffix = @hdpqa.example.com
activeDirectoryRealm.groupRolesMap =
"CN=hdpdv_admin,DC=hdpqa,DC=example,DC=com":"admin"
activeDirectoryRealm.authorizationCachingEnabled = true
sessionManager = org.apache.shiro.web.session.mgt.DefaultWebSessionManager
cacheManager = org.apache.shiro.cache.MemoryConstrainedCacheManager
securityManager.cacheManager = $cacheManager
securityManager.sessionManager = $sessionManager
securityManager.sessionManager.globalSessionTimeout = 86400000
```


HORTONWORKS

User reports: Can’t see interpreter Page

- Zeppelin has URL based access control enabled
- User does not have the role Or Role incorrectly mapped

```
[main]
activeDirectoryRealm = org.apache.zeppelin.server.ActiveDirectoryGroupRealm
activeDirectoryRealm.systemUsername = XXXXX
activeDirectoryRealm.systemPassword = XXXXXXXXXXXXXXXXX
activeDirectoryRealm.searchBase = DC=hdpqa,DC=Example,DC=com
activeDirectoryRealm.url = ldap://hdpqa.example.com:389
activeDirectoryRealm.principalSuffix = @hdpqa.example.com
activeDirectoryRealm.groupRolesMap =
"CN=hdpdv_admin,DC=hdpqa,DC=example,DC=com":"admin"
activeDirectoryRealm.authorizationCachingEnabled = true
sessionManager = org.apache.shiro.web.session.mgt.DefaultWebSessionManager
cacheManager = org.apache.shiro.cache.MemoryConstrainedCacheManager
securityManager.cacheManager = $cacheManager
securityManager.sessionManager = $sessionManager
securityManager.sessionManager.globalSessionTimeout = 86400000
```

User reports: Livy interpreter fails to run with access error

- Ensure Livy has ability to proxy user

```
Edit HDFS core-site.xml via Ambari:
<property>
<name>hadoop.proxyuser.livy_qa.groups</name>
<value>*</value>
</property>
<property>
<name>hadoop.proxyuser.livy_qa.hosts</name>
<value>*</value>
</property>
```

- Ensure Livy has Impersonation enabled

```
In /etc/livy/conf/livy.conf
livy.impersonation.enabled true
```


HORTONWORKS

Apache Zeppelin: Credentials

Credentials in Zeppelin

- LDAP/AD account
  - Zeppelin leverages Hadoop Credential API
- Interpreter Credentials
  - Not solved yet

This is still an open issue

HORTONWORKS

Thank You

Vinay Shukla

@neomythos

HORTONWORKS