Hadoop Security

Building a fence around your Hadoop cluster

Lars Francke

June 12, 2017

Berlin Buzzwords 2017

Introduction

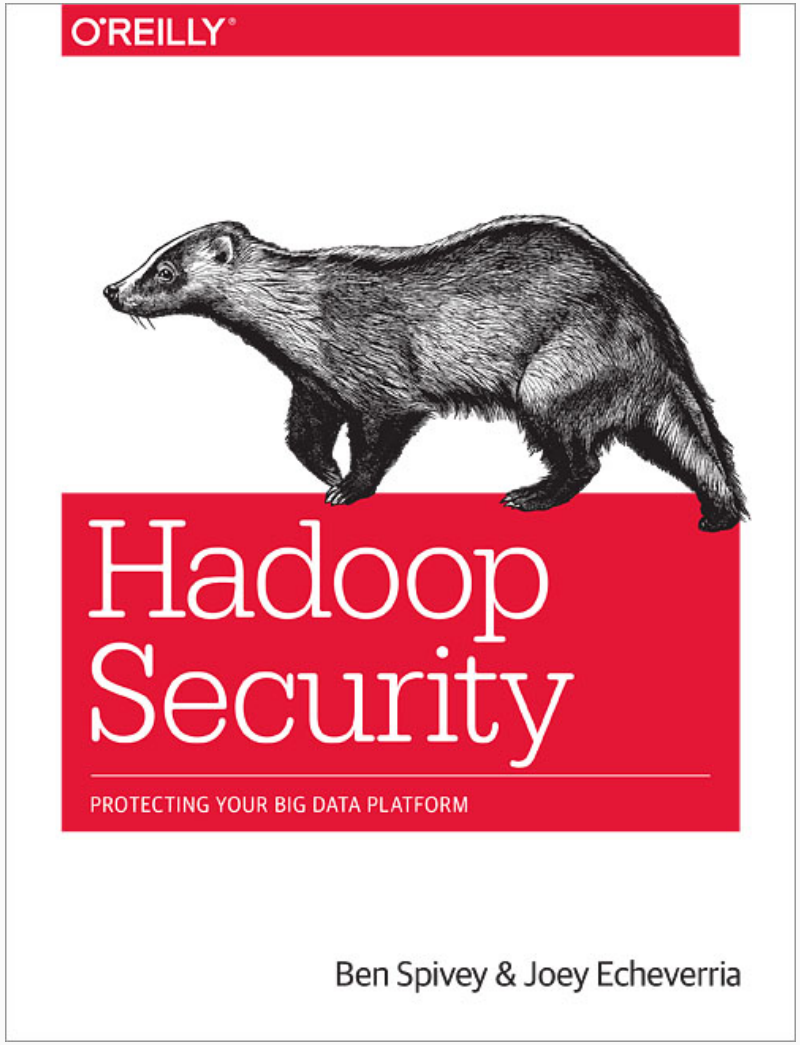
About me - Lars Francke

- Partner & Co-Founder at OpenCore
- Before that: EMEA Hadoop Consultant
- Hadoop since 2008/2009
- Apache Committer: Hive, ORC
- Contact:
  - lars.francke@opencore.com
  - @lars_francke

Overview

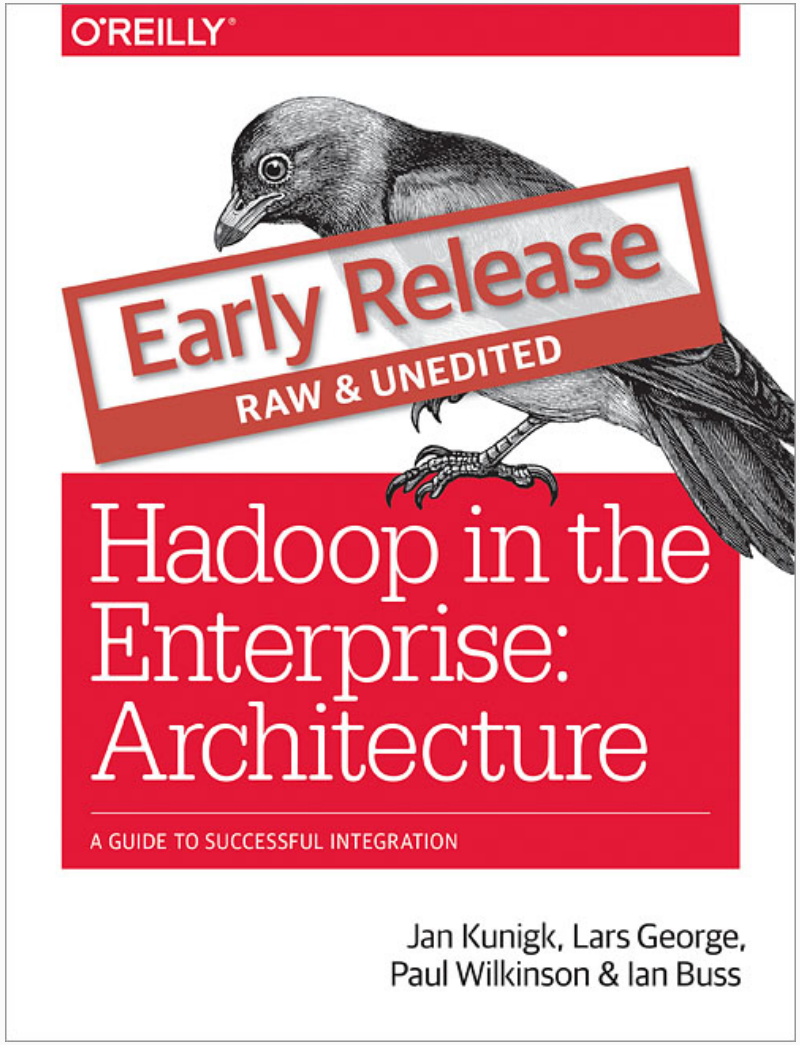
Overview - Be Warned

The Book



Source: oreilly.com

Other Books



Source: oreilly.com

Other Sources

- Kerberos: The Definitive Guide[1]
- Kerberos (German)[2]
- Active Directory, 5th Edition[3]
- Hadoop and Kerberos: The Madness beyond the Gate[4]
- HBase: The Definitive Guide, 2nd Edition[5]
- Bulletproof SSL and TLS[6]

[1] http://shop.oreilly.com/product/9780596004033.do
[2] http://www.kerberos-buch.de/
[3] http://shop.oreilly.com/product/0636920023913.do
[4] https://www.gitbook.com/book/steveloughran/kerberos_and_hadoop/details
[5] http://shop.oreilly.com/product/0636920033943.do
[6] https://www.feistyduck.com/books/bulletproof-ssl-and-tls/

Project Management

How is a typical project structured and where does *Hadoop security* come into play?

Project Management

- Idea/Initiation

Project Management

- Idea/Initiation
- Planning/Design

Project Management

- Idea/Initiation
- Planning/Design
- Execution/Implementation

Project Management

- Idea/Initiation
- Planning/Design
- Execution/Implementation
- “Production”

You can not begin thinking about Security *too early*!

Next steps...

Let’s dig into the details of each step in a project

Project Planning

What To Think About?

- What am I going to use the cluster for?

What To Think About?

- What am I going to use the cluster for?
- Which tools am I going to need to achieve the use-cases?

What To Think About?

- What am I going to use the cluster for?
- Which tools am I going to need to achieve the use-cases?
- What kind of data am I going to ingest, store or process?

What To Think About?

- What am I going to use the cluster for?
- Which tools am I going to need to achieve the use-cases?
- What kind of data am I going to ingest, store or process?
- What kind of corporate guidelines exist that must be followed?

Do Not Assume Anything

Do not *assume* companies/people know what a product is capable of just because they seem confident.

What is part of *Hadoop Security*?

Parts Of A Security Concept

- Authentication of users

Parts Of A Security Concept

- Authentication of users
- Authorization of users

Parts Of A Security Concept

- Authentication of users
- Authorization of users
- Auditing

Parts Of A Security Concept

- Authentication of users
- Authorization of users
- Auditing
- Data Protection: Encryption on-the-wire

Parts Of A Security Concept

- Authentication of users
- Authorization of users
- Auditing
- Data Protection: Encryption on-the-wire
- Data Protection: Encryption at-rest

Project Execution

Kerberos == Hadoop Security?

So, does Hadoop Security mean I need to “enable” Kerberos and I’m done?

Kerberos == Hadoop Security?

So, does Hadoop Security mean I need to “enable” Kerberos and I’m done?

No! Far from it!

Overview

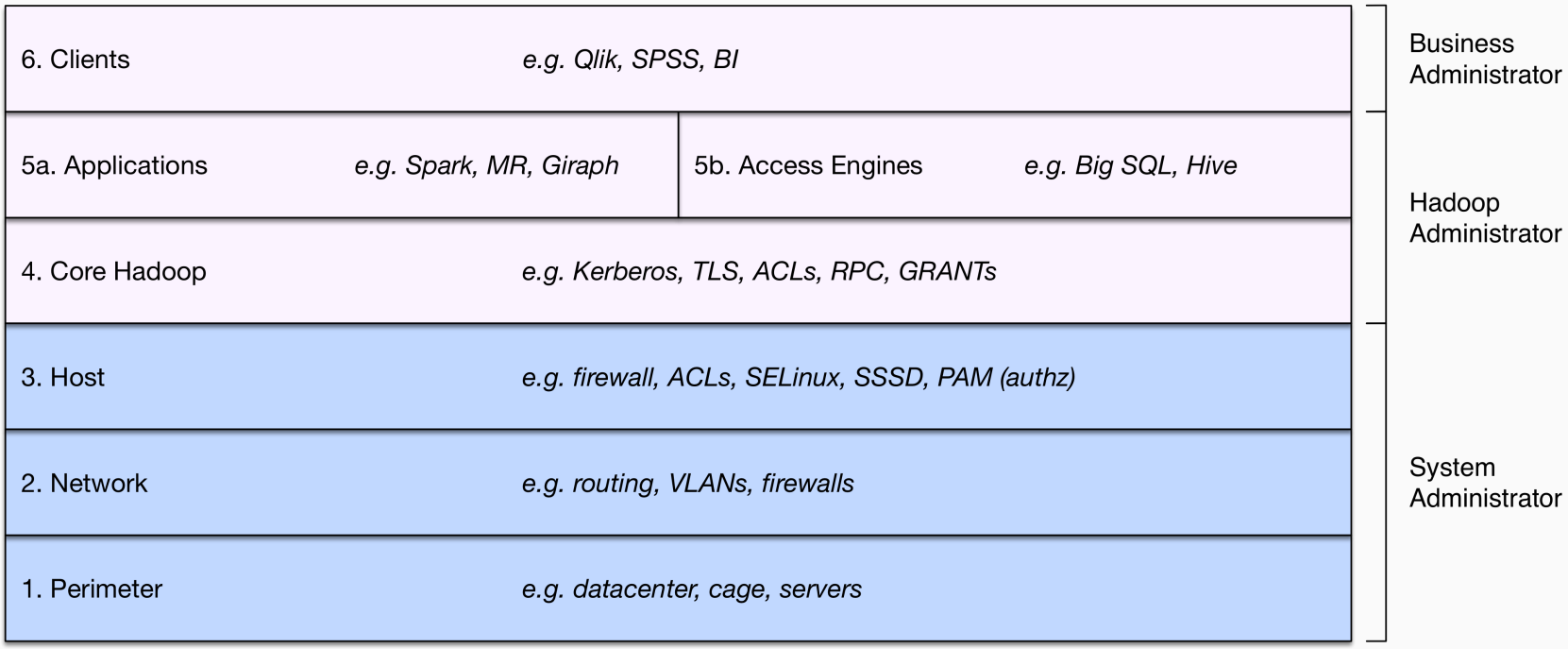
Before you’re able to secure anything...

Overview

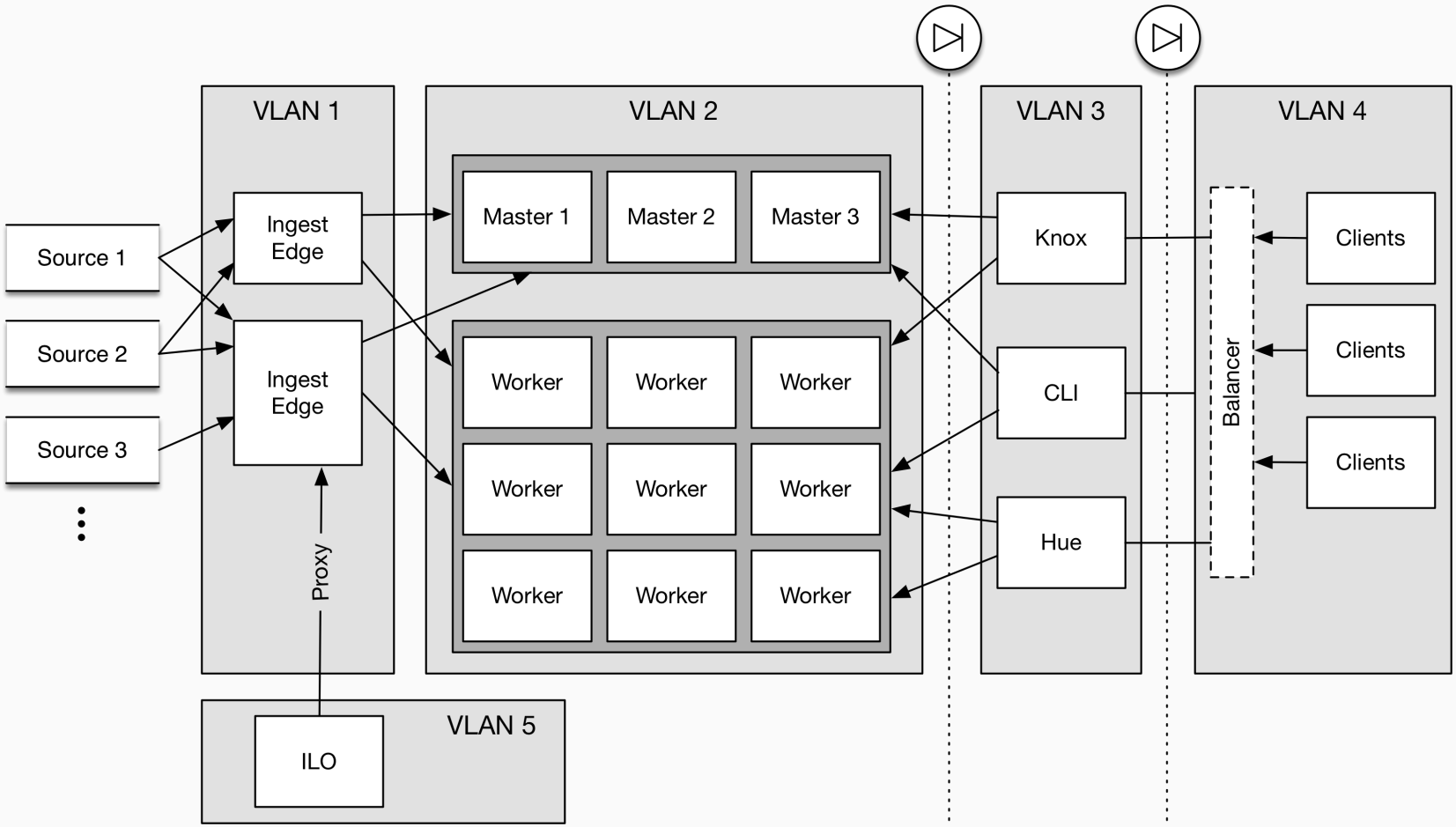
Before you're able to secure anything...

...something must be running that *can* be secured.

Overview - contd.



Network



Host/Operating System

- SELinux
- NTP/Chrony
- Firewalls
- Antivirus
- Proxies

Hadoop Installation

- You usually install a cluster manager first

Hadoop Installation

- You usually install a cluster manager first
- Then you install Agents

Hadoop Installation

- You usually install a cluster manager first
- Then you install Agents
- Followed by a distribution with lots of components...

Authentication

- Core Hadoop requires Kerberos for strong authentication
- Multiple choices on how to implement that:
  - Cluster-local standalone KDC
  - Cluster-local standalone KDC with one-way cross-realm trust to a central KDC
  - Direct integration into a central KDC

Authentication: Cloudera Manager/Ambari Wizards

Source: unsplash.com by Sirotorn Sumpunkulpak

Authentication - contd.

What you need:

- Names/IPs for your KDC
- Supported Encryption Types (see my blog post[7] for more details)
- Firewall must allow all cluster machines to access the KDC
- Potentially a bunch of information to configure `krb5.conf` properly

[7] http://www.opencore.com/blog/2017/3/kerberos-encryption-types

Authentication - contd.

Ideally you want the automatic option, but...

- You need an account in your AD/KDC that is allowed to create other accounts!
- (AD only) You need to talk via LDAPS and make sure that you have all the truststores set up properly
- (AD only) You cannot have multiple SPNs per User
- For SPNEGO you need a `HTTP/<host>` principal, sometimes doesn't match corporate guidelines

Authentication - contd.

There’s a manual option...

Authentication - contd.

Kerberos is only part of the authentication story.

Authentication - contd.

Knox, Cloudera Manager, Ambari, Ranger, Sentry, Hive, Impala etc. all have their own authentication layers which also support LDAP(S) or other mechanisms.

Identity Management

Your users need to *exist* on the workers so that YARN can start jobs using `setuid()`.

Identity Management - contd.

- This is usually done using a tool like SSSD (free) or Centrify (commercial)

Identity Management - contd.

- This is usually done using a tool like SSSD (free) or Centrify (commercial)
- When using Centrify, configure it to *not* create the default HTTP principal for each server

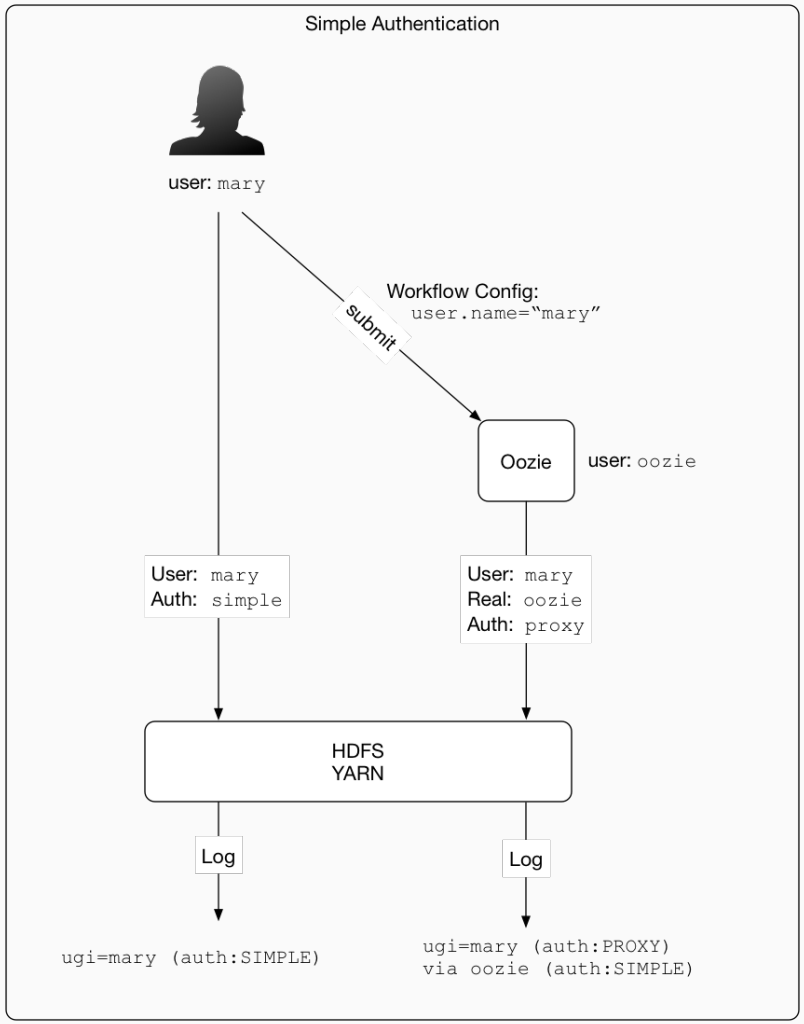
Identity Management - contd.

- This is usually done using a tool like SSSD (free) or Centrify (commercial)
- When using Centrify, configure it to *not* create the default HTTP principal for each server
- This does *not* mean that your users need to be able to SSH into the workers, quite the opposite

Identity Management - contd.

- You need to have the details needed to fetch user & group information from LDAP(S)
- Some solutions require a domain join

Impersonation/Proxy Users



Authorization

- All tools have some form of authorization built-in
- Core Hadoop components have a first line defense: Service Level Authorizations
- Ranger & Sentry promise cross-cutting RBAC functionality

Data Protection: Encryption At-Rest

HDFS Transparent Encryption

All I need is a KMS, right?

Data Protection: Encryption At-Rest - contd.

- KMS should be “close” to your clients and NameNode
- KMS should be separately administered
- AFAIK only Hadoop 3 will allow the / (root) directory to be its own zone and have sub-zones
- Secure access to the backing store & host where KMS runs
- You can use a HSM
- What about Impersonation and things like Knox or HttpFS?

Data Protection: Encryption At Rest - contd.

Now that my data in HDFS is secure I'm good, right?

- Metadata in databases
- Loq & Audit files
- YARN localized stuff
- Temporary files
- (Spark) spill files & cached data

Only Cloudera has a (paid) solution for this: Cloudera NavEncrypt

Data Protection/Authorization: Data Masking

- Static vs. dynamic masking
- Ranger & BigSQL support masking, Sentry does not
- RecordService might be a potential integration point
- 3rd party tools

Data Protection: Encryption On-The-Wire

- Web Interfaces
- REST/Thrift Interfaces
- MapReduce Shuffle
- Spark Shuffle (only in 2.1+)
- Server to Agent communication
- Ingest/Egress tools (Flume, Informatica, Kafka, ...)
- RPC
- HDFS Data traffic
- Traffic from your YARN apps

Data Protection: Encryption On-The-Wire

- 2-way TLS possible?
- Which cipher suites are supported?
- Some tools automatically disable HTTP others don’t!
- Missing documentation all around

Auditing & Logging

- No documentation (on auditing events and lots of other things)
- Ranger aggregates audit logs, async by default
- Cloudera has Navigator
- No integrity protection
- Need to protect against admins

3rd Party Tools

- Not a single product has good documentation
- How do they access the cluster?
- Do they themselves authenticate and authorize users? How?
- Auditing/Logging?

Production

What now?

So you've finished the project and handed it over to the operations team...what now?

Keep it running

- Plan regular updates to your OS
- Use a Configuration Management tool (e.g. Ansible)
- Plan *regular* updates to your distribution!

Monitoring

All our logging & auditing is irrelevant if you're not monitoring & alerting on those logs.

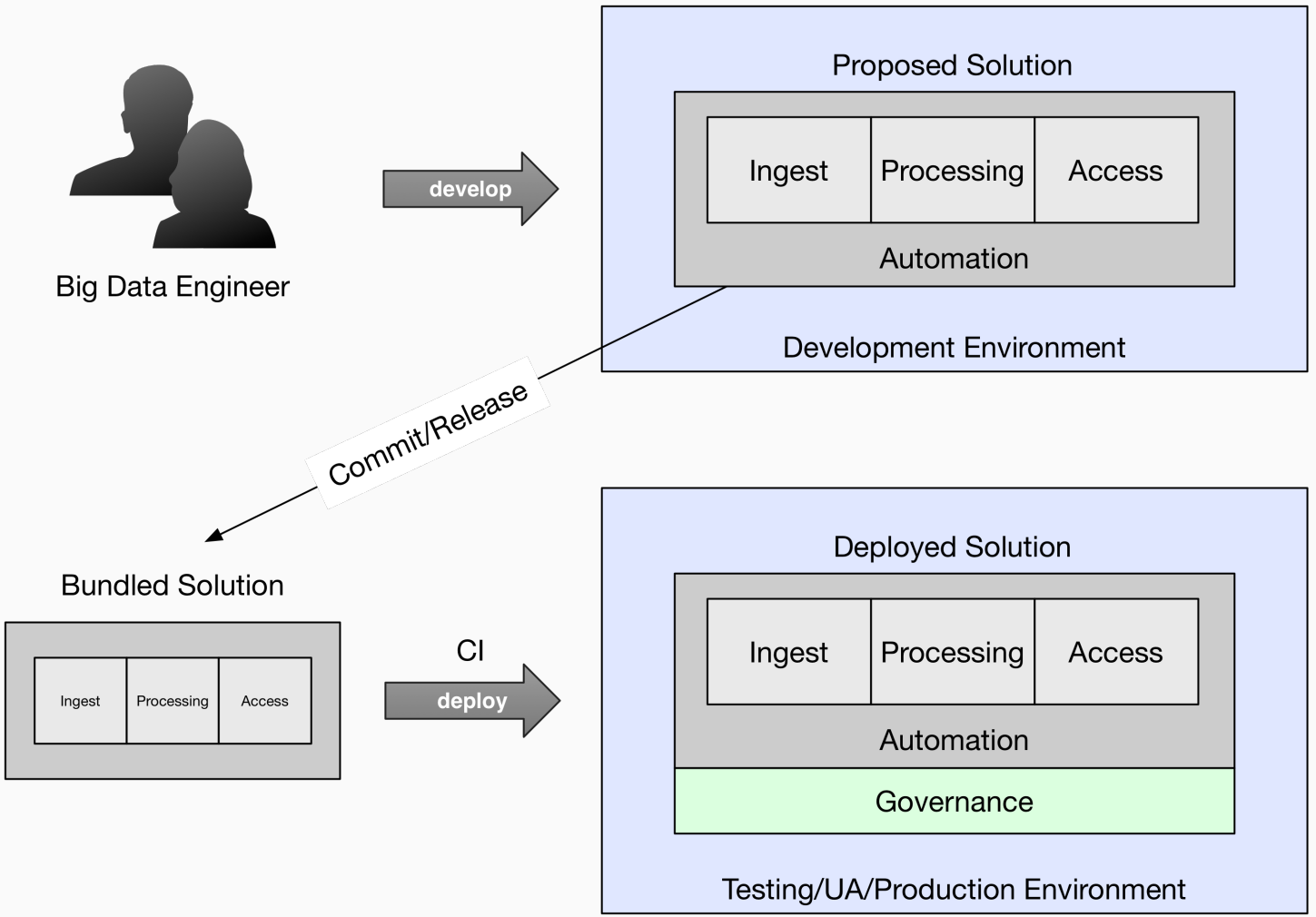
Environments

You need a place where you can test upgrades.

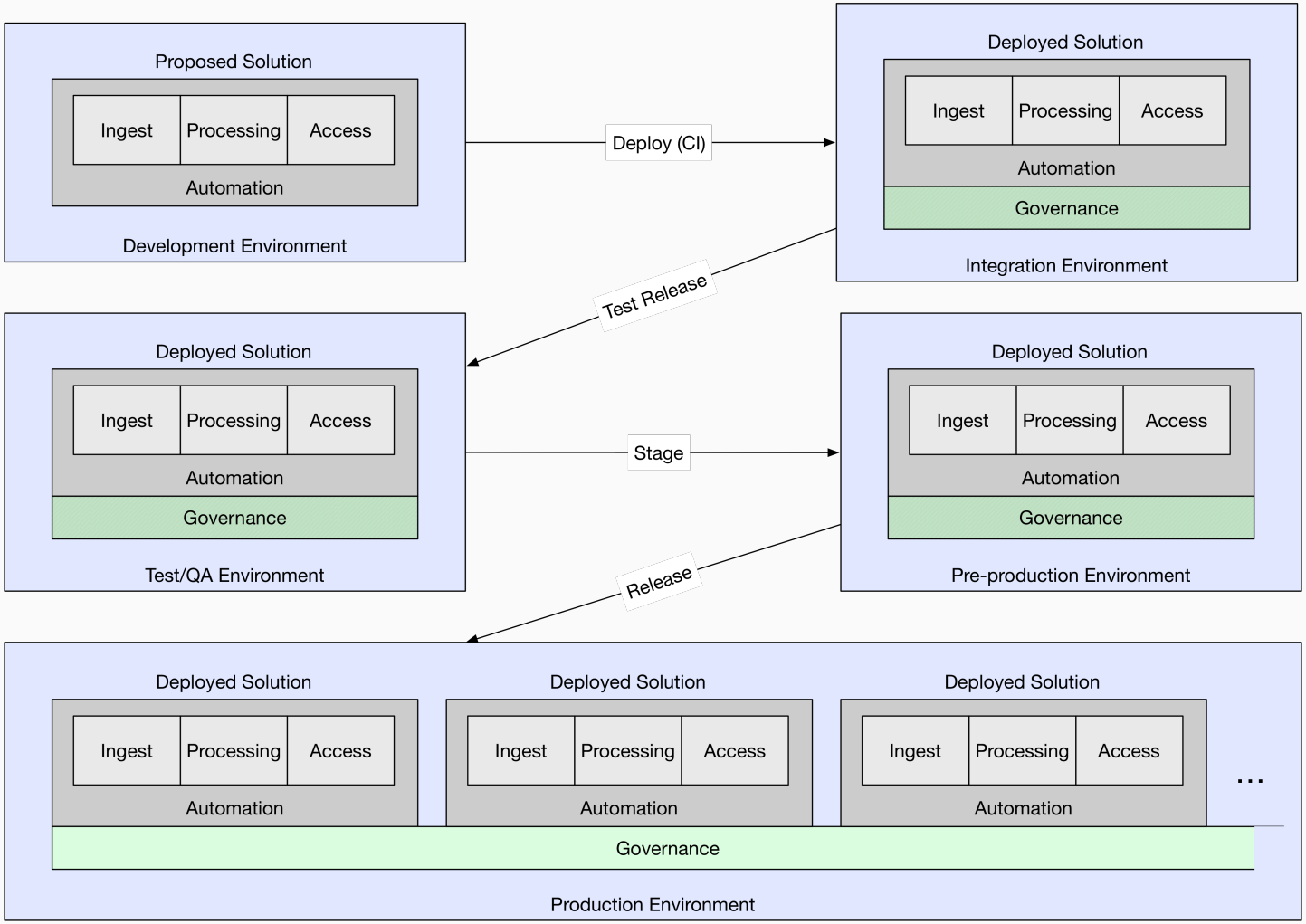
You might also need a place for backups.

Devs would like a cluster as well...

Environments



Environments



User Lifecycle

- What happens when a user leaves your organization?
- Caches
- Ranger Usersync
- Existing sessions
- Shared accounts

Misc

Cloud?

What about the cloud providers?

Distributions

- Cloudera

Distributions

- Cloudera
- Hortonworks

Distributions

- Cloudera
- Hortonworks
- IBM

Distributions

- Cloudera
- Hortonworks
- IBM
- Microsoft HDInsight

Thank You!

Thank you for listening!

Questions?

Contact me at:

- lars.francke@opencore.com
- @lars_francke

Visit us at opencore.com

And if this stuff interests you: We're looking to expand our team!