Securing Hadoop in an Enterprise Context

Hellmar Becker, Senior IT Specialist

Apache: Big Data conference

Budapest, September 29, 2015

ING

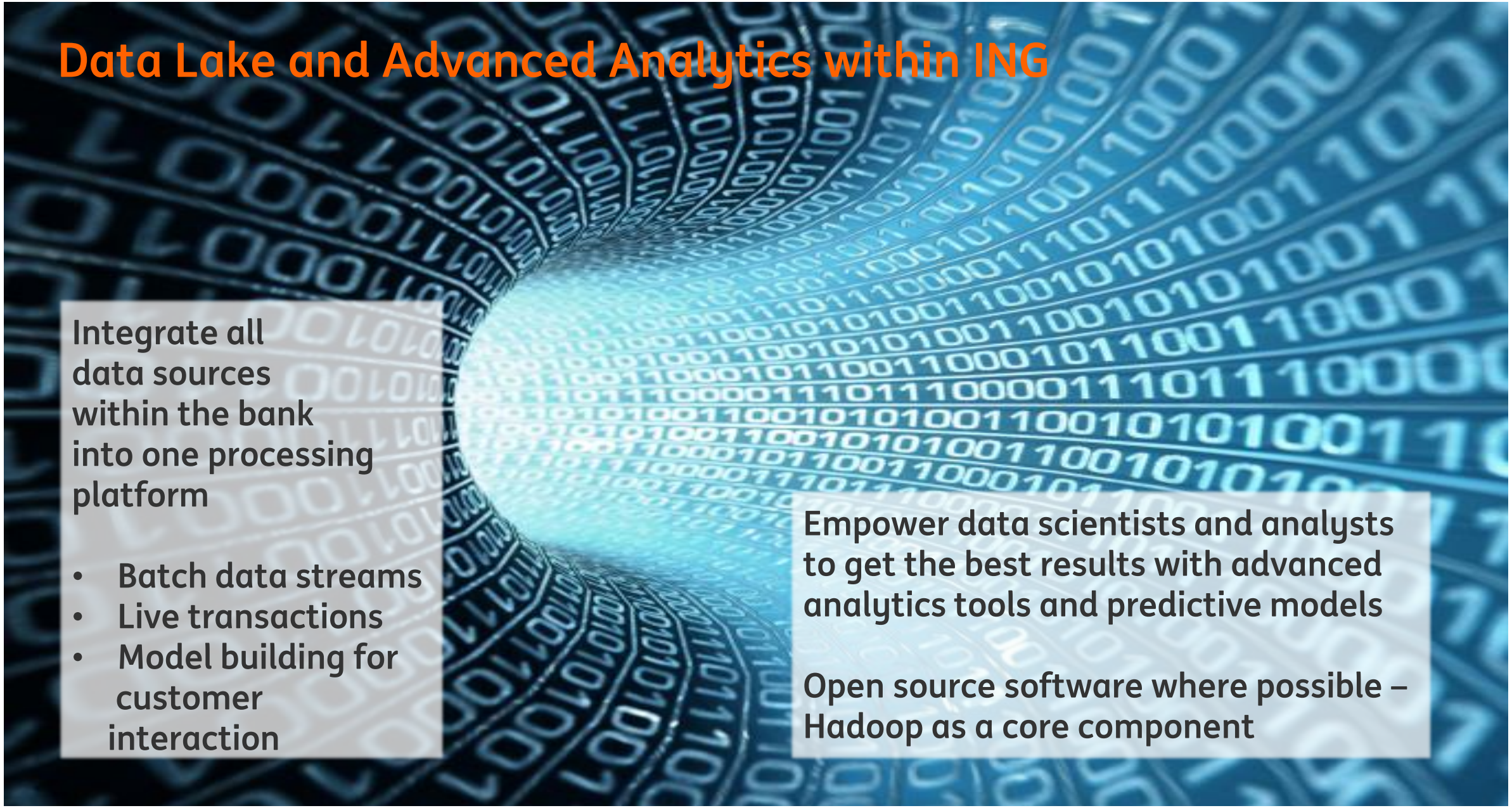
Who am I?

Securing Hadoop in an Enterprise Context

1. The Challenge
2. Excursion: Hadoop Usage Patterns
3. Aspects of Security
4. Analytic Clusters: “Sandbox” Model
5. Securing HDFS Environments That Do Automated Processing
6. Connecting to the Enterprise Directory
7. Further Aspects
8. Questions

1. The Challenge

Data Lake and Advanced Analytics within ING



Integrate all data sources within the bank into one processing platform

- Batch data streams
- Live transactions
- Model building for customer interaction

Empower data scientists and analysts to get the best results with advanced analytics tools and predictive models

Open source software where possible – Hadoop as a core component

ING

Risks

- Data loss
- Privacy breach
- System intrusion

Possible consequences

Legal consequences
Loss of reputation
Financial loss

Hadoop "out of the box" does not have any security model switched on

Hadoop user model:

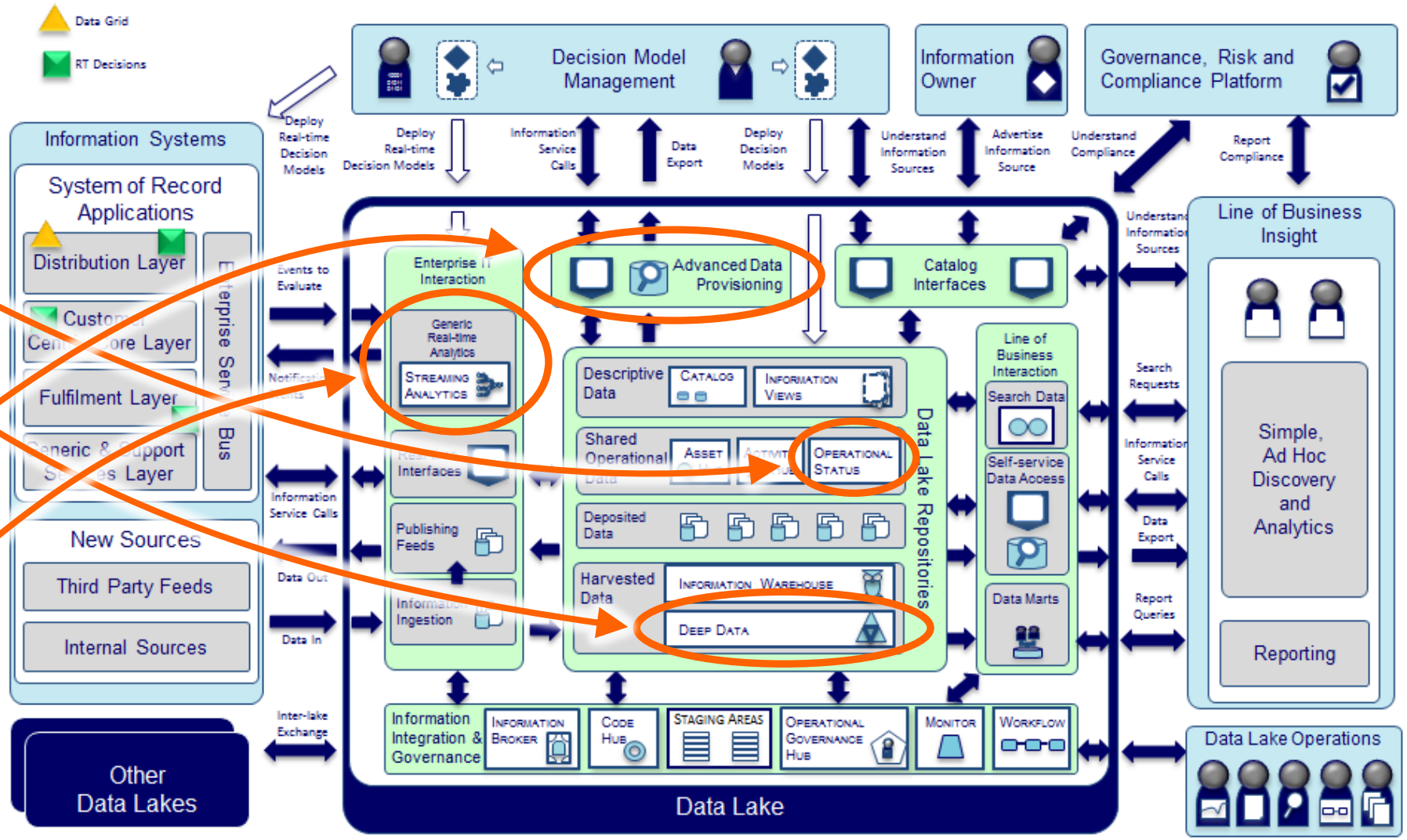
- A user name is just an alphanumeric string
- So is a group name
- They do not have to match entities in the OS
- Via REST API anybody could in theory read/write HDFS

2. Excursion: Hadoop Usage Patterns

ING

Hadoop Usage Patterns

1. File Storage
2. Deep Data
3. Analytical Hadoop
4. (Real Time)



ING

Hadoop Usage Patterns: Characteristics

Topics	Analytical Hadoop	Deep Data	File Storage
User Access	Named	Non Personal Accounts	Non Personal Accounts
Capacity mgmt.	Small disk space	Large disks space	Large disks space
Resource mgmt.	High CPU & memory	Med CPU & memory	Low CPU & memory
Confidentiality Integrity Availability – rating	C based on use case, IA-low	C static/data driven, IA-high	C static/data driven, IA-high
Flexibility	High	Low	Low
Tooling outside Hadoop	High & user driven	Low & life cycle driven	Low & life cycle driven
Disaster recovery & High Availability	Low	High	High
Predictability of Jobs	Ad hoc	Scheduled	None
Data	Subset relevant for use case	All	All
Lineage	Irrelevant	Relevant	Relevant
Descriptive metadata	Relevant	Relevant	Relevant
Develop Test Acceptance Production	Develop (Test)	Test Acceptance Production	Test Acceptance Production

3. Aspects of Security

Aspects of Security

Technical: Rings of Defense

- Perimeter Level Security
- Application Level Authentication and Authorization
- OS Security
- Data Protection

See also: http://www.slideshare.net/vinnies12/hadoop-security-today-tomorrow-apache-knox

Conceptual: Five Pillars of Security

- **Administration**
- **Authentication**
- **Authorization**
- **Auditing**
- Data Protection

See also: http://hortonworks.com/hdp/security/

ING

4. Analytic Clusters: “Sandbox” Model

Approach A: “Sandbox”

- Strong perimeter security
- Ideally "air gapped"
- Practical: allow access only through a terminal service (Citrix, VNC)

Pro:

- Easy to implement
- No changes to internal settings

Con:

- Even legitimate data transfers are difficult
- Not suitable for automated batch processing
- Software updates only through manually maintained mirror

Used in exploratory environments (pattern 3)

5. Securing HDFS Environments That Do Automated Processing

ING

Administration

- General goal: Zero Touch deployment
- Automatic synchronization with enterprise directory
- Ranger UI is only used for incidents

Authentication

- Kerberos
- Question of one KDC per Cluster? (Yes)
- Connecting to enterprise directory (next chapter)
- Keep the Kerberos principals (Hadoop users) completely separate from OS users

ING

Authorization

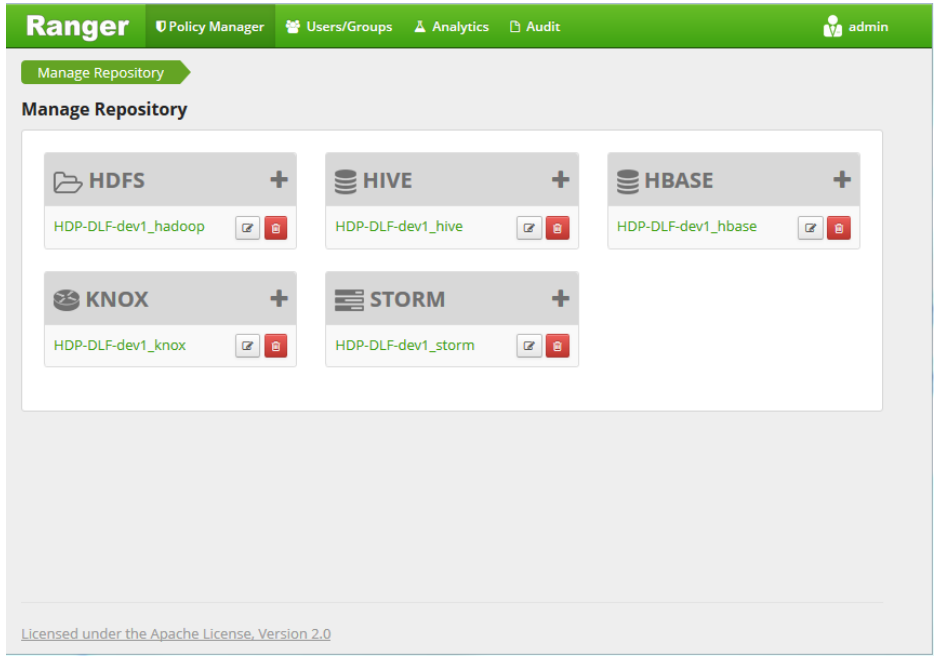
Simplest approach: HDFS ACLs

BUT:

- No easy to use GUI
- Difficult to maintain overview
- Only for HDFS, does not handle other components

```
> hdfs dfs -setfacl -m group:execs:r-- /sales-data

> hdfs dfs -getfacl /sales-data
# file: /sales-data
# owner: bruce
# group: sales
user::rw-
group::r--
group:execs:r--
mask::r--
other::---
```



Better: Unified rights management with Ranger

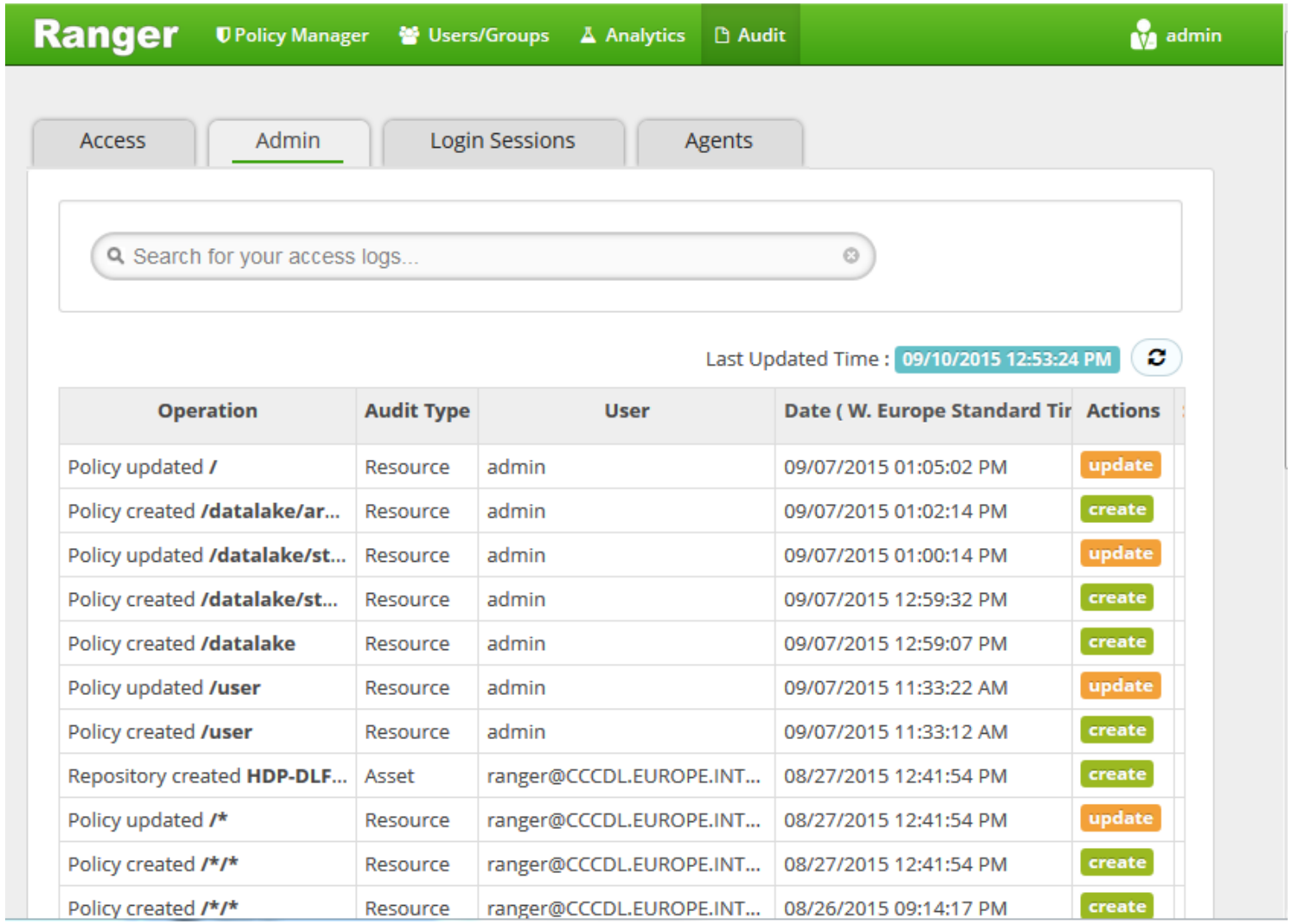
- Service principals will be directly made known to Ranger; PA's rights are assigned only based on groups
- Groups and users are synced with AD. See below for details

- Note: Be aware that Ranger can not take away privileges that were granted on a lower level
- HDFS permissions and ACLs override Ranger
- Make sure these access paths are locked down

ING

Auditing

- Ranger standard auditing
- More testing required: Is audit logging to a database good enough/fast enough?

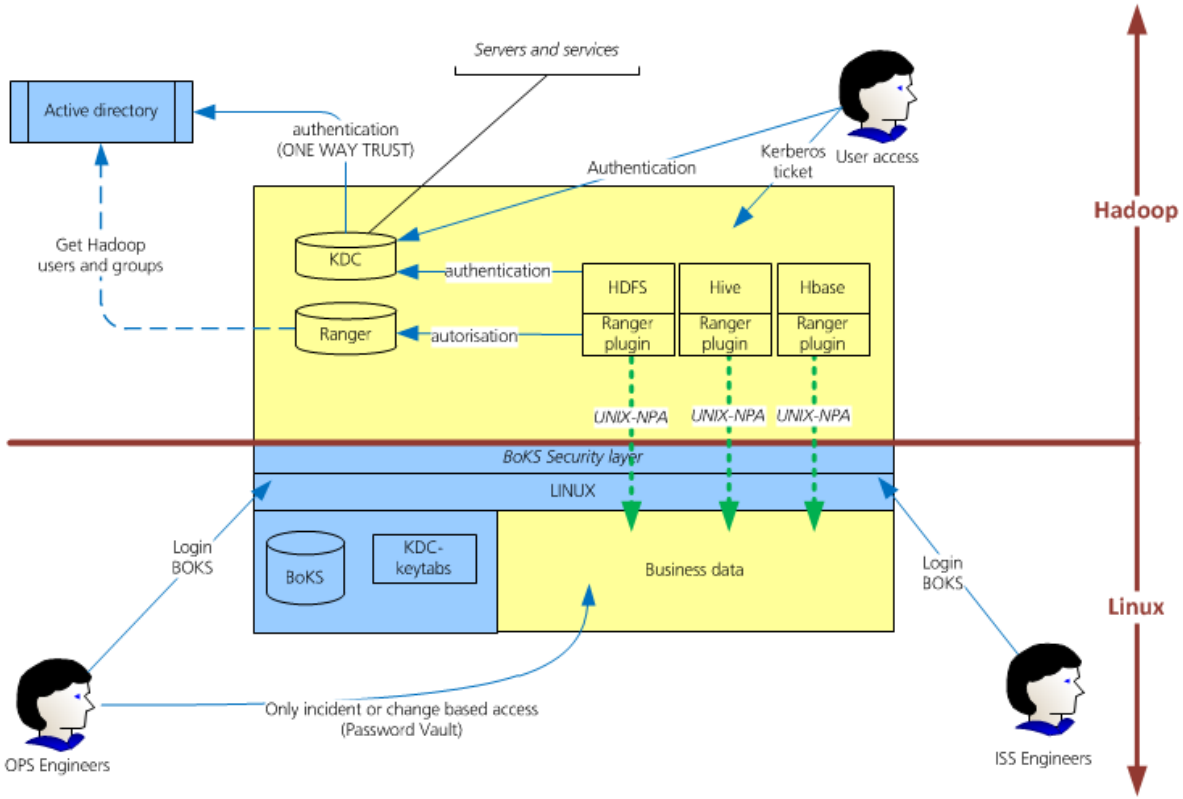


Operation	Audit Type	User	Date (W. Europe Standard Tir	Actions
Policy updated **/**	Resource	admin	09/07/2015 01:05:02 PM	update
Policy created **/datalake/ar...**	Resource	admin	09/07/2015 01:02:14 PM	create
Policy updated **/datalake/st...**	Resource	admin	09/07/2015 01:00:14 PM	update
Policy created **/datalake/st...**	Resource	admin	09/07/2015 12:59:32 PM	create
Policy created **/datalake**	Resource	admin	09/07/2015 12:59:07 PM	create
Policy updated **/user**	Resource	admin	09/07/2015 11:33:22 AM	update
Policy created **/user**	Resource	admin	09/07/2015 11:33:12 AM	create
Repository created **HDP-DLF...**	Asset	ranger@CCCDL.EUROPE.INT...	08/27/2015 12:41:54 PM	create
Policy updated **/***	Resource	ranger@CCCDL.EUROPE.INT...	08/27/2015 12:41:54 PM	update
Policy created **/*/***	Resource	ranger@CCCDL.EUROPE.INT...	08/27/2015 12:41:54 PM	create
Policy created **/*/***	Resource	ranger@CCCDL.EUROPE.INT...	08/26/2015 09:14:17 PM	create

6. Connecting to the Enterprise Directory

Separation of administrative duties

- Personal users in corporate Active Directory, NPAs in cluster KDC
- One way realm trust



Specific challenges

- Historically, Windows and Linux are different worlds
- Need to work in interdisciplinary teams
- Educate AD experts on the details of Kerberos realm trust
- Still to be solved: YARN containers need to run as a OS user that matches the HDFS user name
  - AD and Linux LDAP use different user keys
  - Currently, some teams use workarounds for this (manually maintenance required)

Security roles for personal accounts

- Maintained in HR database/tools
- More interdisciplinary cooperation required!
- Need to map abstract "business roles" (function descriptions) to "technical roles" (sets of privileges)
- HR database maintainers have to update this, it will be reflected in AD
- In LDAP, these technical roles appear as groups

Synchronizing users and roles from Active Directory

- Ranger's *uxugsync* process queries Active Directory through LDAP protocol

- **Ranger 0.4:** Reads all users, then determines their group affiliation
- More than 50,000 employees in ING Group
- Need to limit the load on LDAP server!

- **Ranger 0.5:** Group driven query - still not optimal because it uses attribute filters
- Most efficient LDAP query is either by a single DN (Distinguished Name), or by container (query base DN).
- But we cannot use containers because of enterprise policy
- **Solution**: custom Python script that queries LDAP hierarchically
- One “supergroup” is picked by DN
- The members of the “supergroup” are all LDAP groups that have Hadoop related privileges
- Query all these groups, again by DN
- Examine the members of each group (personal users)
- Make the user-group relationships known to Ranger via REST call

ING

7. Further Aspects

Securing the Non-Kerberos/Ranger Components

- Use LDAP to authenticate in Ambari, Hue
- Note: Our current setup connects Ambari to Unix LDAP, which is not in sync with AD

Securing the Perimeter

- Knox
- Reverse proxy

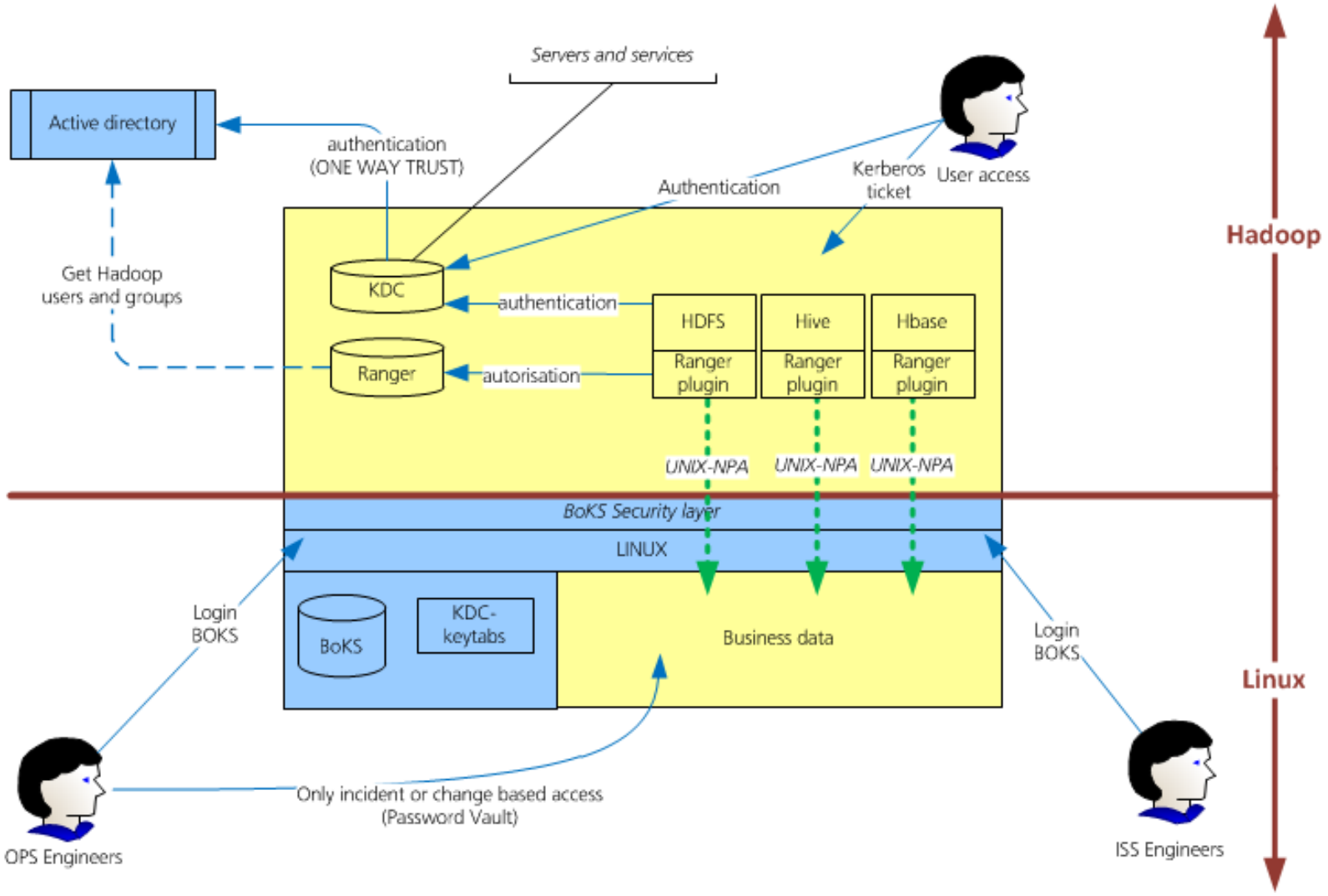
Securing Platform Components

- A good HDFS security model takes care of much that follows
- Considerations for database-like processing (Hive, Hbase): Column or file based security models, can't have both

8. Questions

Attributions

- Hellmar in Nîmes / With Python in Mindanao, by the author
- Domtoren in het oranje licht by helena_is_here is licensed under CC BY 2.0
- Data Pipeline, ING OIB Image Bank
- Storm surge by David Baird is licensed under CC BY-SA 2.0; cropped by me
- System Lock by Yuri Samoilov is licensed under CC BY 2.0; cropped by me
- Safe by Rob Pongsajapan is licensed under CC BY 2.0; cropped by me
- Hercules and Cerberus by The Los Angeles County Museum of Art is Public Domain

Backup

Security Model