HADOOP UNDER ATTACK

SECURING DATA IN A BANKING DOMAIN
WHOAMI

Federico Leven

- @ ReactoData
- Big Data + Open Source from 2012
- Big Data Meetup coordinator
  (http://www.iaar.site), speaker ...
- federico@reactodata.net
- Web : http://www.reactodata.net
- Twitter: @reactodata
- Linkedin : https://www.linkedin.com/in/federicoleven/
We are a startup based in Buenos and Poland, providing Big Data + Cloud solutions based on Open Source and proprietary software and Hadoop consultancy.

- Big Data and Hadoop applications development
- Machine Learning
- Cloud
- UX/UI and Mobile Apps for Big Data platforms
- Hadoop Consultancy
Agenda

- The Challenge: Best Practices + Regulations
- How to do it in Hadoop
- End-to-End Secured Architecture
- What can go wrong?
- References
- Conclusion & Questions
The Challenge
Best Practices
and regulations
The Challenge : Data Security

The set of preventive, detective and corrective measures to protect the integrity, confidentiality and availability of the data.

CAAIN

• CONFIDENCIALITY
• AVAILABILITY
• AUTHENTICITY
• INTEGRITY

- NON-REPUDIATION

❑ ACCOUNTABILITY / AUDITING
❑ TRACEABILITY
C(A)AIN

- **CONFIDENCIALITY**: Data is not made available or disclosed to unauthorized parties.

- **AVAILABILITY**: Data is available when is needed.

- **AUTHENTICITY**: Data source identity is verifiable.

- **INTEGRITY**: Data is accurate and complete over its entire lifecycle.

- **NON-REPUDIATION**: Parties of a data transaction cannot deny having received/sent the data.
The Challenge: Threats in financial and banking domain

Emerging Technologies Challenges
- Botnet
- IoT unsecured devices
- DDoS (Distributed Denial of Service Attack)

Insider Challenges
- Unintentional actions
- Malicious users

Regulation Challenges
- Periodically new and/or stricter regulations
- US Data Protection rules
- EUR: GDPR

Target
- Sensitive data
- Access credentials
The Challenge: Best Practices in banking

**Organization**
- Security Officer
- End User Guidelines
- Access Policies
- Governance
- ...

**Human**
- Employees Awareness
- Training
- ...

**Technological**
- Networking
- Software Updates
- Data Protection
- Auditing
- ...

-
How to do it in Hadoop
From concepts to technology

- AUTHENTICATION: Identify the user.
- AUTHORIZATION: Grant user access to the data.
- PROTECTION: Protect data from being used except by authorized users.
- AVAILABILITY: Make data accessible when needed.
From concepts to technology

C
Authentication
Authorization

A
Availability

A
Authentication

N
Protection
Availability

I

• Kerberos
• LDAP
• Sentry
• HBase ACLs
+}

• Hadoop
• HA (HDFS, HBASE ...)

• Kerberos
• LDAP

• Encryption
(Motion & Rest)
• Redaction
• Hadoop/HA

Auditing
Traceability

Metadata
Lineage
Log Audit

Cloudera
Navigator
From concepts to technology

- BCRA A6375
- BCRA A6495
- ISO 17799/27001

https://www.cloudera.com/documentation/enterprise/5.14-x/topics/sg_edh_overview.html
What we needed in a banking infrastructure for Hadoop
End-to-End
Secured Architecture
Example Production deployment (CDH 5.13)
Secure data pipeline example

Sources
- Rabbit MQ
- ORACLE

Ingest
- Flume Agent
- Sqoop

HDFS Landing Area
- HDFS

HDFS Business Area
- Impala
- Spark SQL

Web UI
- Spark ETL
- Encrypted Zone

Data redaction custom component
Sqoop SSL ON

ORACLE
S q o o p S S L O N
What can go wrong?
What can go wrong? Some good news and some bad news

- UNSCURE APPLICATIONS WILL NOT WORK ON SECURE ENVIRONMENTS

- Sentry HDFS synchronization does not support Hive Metastore HA (CDH 5.9)

- Sentry HA not supported (CDH 5.9)

- To use CM Kerberos wizard, you need a high level privileges user

- SparkSQL does not respect Sentry permissions (Latest)

- Enabling Sentry turns off Hive impersonation (CDH 5.9)

- Spark Streaming cannot consume from secure Kafka (CDH 5.9)
References
References

✓ http://www.bcra.gob.ar/Pdfs/Texord/t-rmsist.pdf
✓ https://www.cloudera.com/documentation/enterprise/5-9-x/topics/security.html
Thank you!

Questions, suggestions or complaints?

“No Hadoop was harmed in the making of this presentation”