Big problems with big data – Hadoop interfaces security

Jakub Kaluzny

ZeroNights, Moscow, 2015
Sr. IT Security Consultant at SecuRing

• Consulting all phases of development
• penetration tests
• high-risk applications and systems

Researcher

• Hadoop, FOREX, MFP printers, proprietary network protocols
Agenda

- Big data nonsenses
- Crash course on hacking Hadoop installations
- Ways to protect big data environments

Expect some CVEs
Results summary

- no account
- standard user
- admin user

data access

admin privileges
WHAT IS HADOOP?

Know your target
Normal database
Normal database architecture

Still normal database scenario

CWE-xxx: SQL Injection through license plate
Normal database injection points
Normal database

Clear rules

Users  Roles  Data  Model

Clear target
Anecdote

<table>
<thead>
<tr>
<th>user db, a lot of clients</th>
<th>critical banking data, one supplier</th>
</tr>
</thead>
</table>

Only one common table

Q: Why don’t you split it into 2 dbs with a db link?
A: Too much effort and we want to have fast statistics from all data.
What is Hadoop?
Hadoop architecture schema
More on Hadoop

Apache Hadoop Ecosystem

- Ambari
  - Provisioning, Managing and Monitoring Hadoop Clusters

- YARN Map Reduce v2
  - Distributed Processing Framework

- HDFS
  - Hadoop Distributed File System

- Flume
  - Log Collector

- Zookeeper
  - Coordination

- Oozie
  - Workflow

- Pig
  - Scripting

- Mahout
  - Machine Learning

- R Connectors
  - Statistics

- Hive
  - SQL Query

- Sqoop
  - Data Exchange

- Apache HBase
  - Column Store
Hadoop injection points
Hadoop scenario

https://www.flickr.com/photos/mattimattila/8349565473

http://bigdataanalyticsnews.com/tag/hortonworks/

https://en.wikipedia.org/wiki/Moneygami
What is a lot of data?

21 PB of storage in a single HDFS cluster
2000 machines
12 TB per machine (a few machines have 24 TB each)
1200 machines with 8 cores each + 800 machines with 16 cores each
32 GB of RAM per machine
15 map-reduce tasks per machine
What is a lot of data?

Our latest assessment:

• 32 machines, 8 cores each
• 24TB per machine
• 64 GB of RAM per machine
• Almost 1 PB disk space and 2TB of RAM

http://mrrobot.wikia.com/wiki/E_Corp
Attacker perspective

https://plus.google.com/+Magiccardtrickszonetips
RISK ANALYSIS

Know your threats
Risk analysis

Who  How  What
Who?

Business perspective: competitor, script-kiddies, APT

Technical perspective:

- **External attacker**
  - Anonymous
  - Ex-employee

- **Insider**
  - Exployee (with some rights in Hadoop): user, admin
  - Infected machine, APT
Risk analysis

Who  How  What
Full compromise

Online banking owned by single attacker
Data safety vs. data security

Security Budget

About 2/3 used to protect the network

Data Protection

Less than 1/3 used to directly protect data and intellectual property

Source: CISO Market Pulse Survey
Q: *What will be stored?*  
A: „*We do not know what data will be stored!*“

**Typical bank scenario**

- All transaction data
- All sales data
- All client data

**Bigdata analytic says:** „*People who bought a dashcam are more likely to take a loan for a new car in the next month*“
For what? Data theft

Forbes / Tech

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

How Facebook knows when you'll get divorced (even before you do)

Facebook knows who your romantic partner is, even if you keep that information private, and can even predict if the relationship will last.
Other

- Privilege escalation
  - Authentication bypass
- Abuse
  - DoS
  - Data tampering
Risk analysis

Who  How  What
How?

WHAT HADOOP REALLY IS

under sales-magic-cloud-big-data cover
Typical architecture

http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview
Hadoop injection points

Differ much amongst distros
INTERFACES
Interfaces

- Distros specifics
- User ifaces
- Admin ifaces
- external ifaces

Hadoop
OUR STORY WITH BIG DATA ASSESSMENT
a.k.a. crash course on hacking big data environments
Interfaces

- Hadoop
- Distros specifics
- User ifaces
- Admin ifaces
- external ifaces
USER INTERFACES

for employees and applications
User interfaces

- Distros specifics
- Admin ifaces
- External ifaces
- User ifaces
User interfaces

Apache Hue
- Pig, Hive, Impala, Hbase, Zookeeper, Mahout, Oozie

Other
- Tez, Solr, Slider, Spark, Phoenix, Accumulo, Storm
Is Hue an internal interface?

http://9gag.com/gag/awrwVL1/hue-hue-hue
Apache Hue overview

Sample: Salary growth

```
SELECT s07.description, s07.salary, s08.salary - s07.salary
FROM sample_07 s07 JOIN sample_08 s08
ON (s07.code = s08.code)
WHERE s07.salary < s08.salary
ORDER BY s08.salary - s07.salary DESC
LIMIT 20
```
Apache Hue DOM XSS

Introducing Hue

Hue is a browser-based environment that enables you to interact with a Hadoop cluster. Hue includes applications run in a Web browser and require no client installation.

Hue Architecture

Hue applications run in a Web browser and require no client installation.

The following figure illustrates how Hue works. Hue Server is a “container” web application that sits in between your HDP installation and the browser. It hosts all the Hue web applications and communicates with various servers that HDP components.

```
var _anchor = $('a[name=""] + decodeURIComponent(window.location.hash.substring(1)) + "]"').last();

Payload: URL/help/#<img src="x" onerror="alert(1)">
```
Apache Hue attack scenario

1. Target old Hadoop installation (with Hue 2.6.1, Django 1.2.3)
2. Target a user with access to Hue
3. Send him XSS
4. Get access to all Hadoop data designated for the user
Default configurations sucks

X-Frame-Options: ALLOWALL

You're seeing this error because you have DEBUG = True in your Django settings file. Change that to False, and Django will display a standard 500 page.
ADMIN INTERFACES

for admins and maintenance
Admin interfaces

- Distros specifics
- User ifaces
- Hadoop
- Admin ifaces
- external ifaces
Admin interfaces

Apache Ambari
- Provisioning, monitoring

Apache Ranger
- Security: authorization, authentication, auditing, data encryption, administration

Other
- Knox, Cloudbreak, Zookeeper, Falcon, Atlas, Sqoop, Flume, Kafka
Trochę o Ambari

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wizard-driven interface</td>
<td>Facilitates installation of Hadoop across any number of hosts</td>
</tr>
<tr>
<td>API-driven installations</td>
<td>Ambari Blueprints for automated provisioning</td>
</tr>
<tr>
<td>Granular service control</td>
<td>Precise management of Hadoop services and component lifecycles</td>
</tr>
<tr>
<td>Configuration change history</td>
<td>Ongoing management of Hadoop service configurations</td>
</tr>
<tr>
<td>RESTful APIs</td>
<td>Enables integration with enterprise systems</td>
</tr>
<tr>
<td>Extensible framework</td>
<td>Brings custom services under management via Ambari Stacks</td>
</tr>
<tr>
<td>Customizable user interface</td>
<td>Develop innovative user experiences via Ambari Views Framework</td>
</tr>
<tr>
<td>User Views</td>
<td>Advanced capabilities for cluster optimization and tuning for Hadoop DevOps</td>
</tr>
</tbody>
</table>

http://www.slideshare.net/hortonworks/ambari-using-a-local-repository?next_slideshow=1
Apache Ambari

Architecture

Ambari Server
- REST API for integration
- Cluster Configurations
- Database (PostgreSQL)
- Python Agents (puppet, python)
- Metrics (ganglia, nagios, jmx)
- Alerts
- Mirroring

Ambari Web
- REST API 100% REST
- Web Client

Request Dispatcher
- Orchestrator
- SPI

Auth Provider
- Configurable Auth Provider
- RDBMS
- AD/LDAP

Pluggable

http://www.slideshare.net/hortonworks/ambari-using-a-local-repository?next_slideshow=1
Is Ambari an internal interface?

[Image of the Shodan website search results for 'ambari']

http://knowyourmeme.com/memes/facepalm
Apache Ambari

- Standard users can sign into Ambari (WHY?)
- Low hanging fruits: directory listing by default, no cookie flags, no CSRF protection
- Interesting proxy script ->
Apache Ambari REST API proxy

Standard request:

limit=1&secondaryFilter=tez:true&_=1424180016625

Tampered request (logs accessible only from DMZ):

/proxy?url=http://google.com
### Directory: /logs/

<table>
<thead>
<tr>
<th>File Path</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>.hadoop-mapreduce.jobsummary.log</td>
<td>137797 bytes</td>
<td>Jan 22, 2015 6:18:54 PM</td>
</tr>
<tr>
<td>yarn-yarn-historyserver.log</td>
<td>3866624 bytes</td>
<td>Feb 16, 2015 11:23:02 AM</td>
</tr>
<tr>
<td>yarn-yarn-historyserver.out</td>
<td>4096 bytes</td>
<td>Feb 14, 2015 2:08:00 PM</td>
</tr>
<tr>
<td>yarn-yarn-historyserver.out.1</td>
<td>828 bytes</td>
<td>Dec 10, 2014 11:51:13 AM</td>
</tr>
<tr>
<td>yarn-yarn-historyserver.out.2</td>
<td>828 bytes</td>
<td>Dec 10, 2014 11:44:31 AM</td>
</tr>
<tr>
<td>yarn-yarn-historyserver.out.3</td>
<td>828 bytes</td>
<td>Dec 10, 2014 10:55:43 AM</td>
</tr>
<tr>
<td>yarn-yarn-resourcemanager.log</td>
<td>19779584 bytes</td>
<td>Feb 16, 2015 11:24:22 AM</td>
</tr>
<tr>
<td>yarn-yarn-resourcemanager.out</td>
<td>171856 bytes</td>
<td>Feb 15, 2015 1:25:50 PM</td>
</tr>
<tr>
<td>yarn-yarn-resourcemanager.out.1</td>
<td>2192 bytes</td>
<td>Dec 10, 2014 12:46:05 PM</td>
</tr>
<tr>
<td>yarn-yarn-resourcemanager.out.2</td>
<td>2086 bytes</td>
<td>Dec 10, 2014 11:46:30 AM</td>
</tr>
<tr>
<td>yarn-yarn-resourcemanager.out.3</td>
<td>2086 bytes</td>
<td>Dec 10, 2014 11:00:48 AM</td>
</tr>
</tbody>
</table>

CVE-2015-1775
Apache Ambari attack scenario

Target old Hadoop installation with Ambari 1.5.0 to 2.0.2

Hijack standard account (or use Hue XSS to perform CSRF)

Log into Ambari, use CVE-2015-1775

Get access to local network (DMZ) – HTTP only

Download logs, exploit other Hadoop servers in DMZ
Admin interfaces

- Distros specifics
- User ifaces
- external ifaces
- Admin ifaces

Hadoop
Apache Ranger overview

Previously: Apache Argus, XA-Secure

Provides central administration for policies, users/groups, analytics and audit data.
Apache Ranger overview

Apache Ranger

- Low hanging fruits: no HTTP hardening, SlowHTTP DoS
- Standard users can log into Ranger but have no permissions
- Interesting function level access control ->
Apache Ranger vulnerabilities
Missing function level access control

- Audit (X)
  - Big Data (X)
  - Admin (V)
  - Login Sessions (X)
    - Session details (X)
    - Show actions (V)
- Users/Group (X)
  - Add new user (V)
  - List (X)
    - List (X)
    - Edit (V)
- Policies/Analytics (V)
  - List (V)
  - Edit (X)
    - Save changes (V)
    - Details (X)
  - Delete (X)

CVE-2015-0266
Apache Ranger attack scenario

- Target an old Hadoop installation (Apache Ranger 0.4 or XA-Secure v. 3.5.001)
- Hijack standard Hadoop account
- Log into Ranger (with low permissions)
- Use CVE-2015-0266 to escalate privileges
- Edit accounts, authorization rules, access policies
Apache Ranger vulnerabilities

https://cwiki.apache.org/confluence/display/RANGER/Apache+Ranger+0.5+-+User+Guide
Apache Ranger XSS through UserAgent

User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0) <script>alert(1);</script>

CVE-2015-0265
Apache Ranger attack scenario

Target an old Hadoop installation (Apache Ranger 0.4 or XA-Secure v. 3.5.001)

Network access to Apache Ranger is necessary (either from the internet or local network)

Log in with any user and password using XSS in UserAgent

You don’t need to escalate privileges, you’re already an admin (after admin opens session tab)

Deploy BEEF or whatsoever (CSRF script) to create users and change policies
Apache Ranger patched

- Affected version: Apache Ranger v 0.4.0, XA Secure v. 3.5.001
- Both vulnerabilities patched in Ranger v 0.5.0
- For a while developers did a self-full-disclosure ->
RANGER-284 in public Jira now

Replace "Agents" with "Plugins" in Ranger Admin UI

<table>
<thead>
<tr>
<th>Details</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Bug</td>
</tr>
<tr>
<td>Priority:</td>
<td>Major</td>
</tr>
<tr>
<td>Affects Version/s:</td>
<td>0.4.0</td>
</tr>
<tr>
<td>Component/s:</td>
<td>None</td>
</tr>
<tr>
<td>Labels:</td>
<td>None</td>
</tr>
</tbody>
</table>

**Status:** RESOLVED  
**Resolution:** Fixed

**People**

- **Assignee:** Gautam Borad
- **Reporter:** Gautam Borad

**Dates**

- **Created:**

---

Review all references to "Agent" in the UI templates and replace them with "Plugin". For Eg:

- Page: Audit-->Agents:
- Search text: "Search for your agents.~"
- Search fields: "Agent Id", "Agent IP"
- Columns: "Agent id", "Agent IP"
RANGER-284 shortly after vendor contact

Gautam Borad updated RANGER-284:

Attachment: RANGER-284-Escape-HTML-before-displaying-to-prevent-.patch

*Sanitize User Data to prevent XSS - Security Vulnerability*

**Key:** RANGER-284
**URL:** https://issues.apache.org/jira/browse/RANGER-284
**Project:** Ranger
**Issue Type:** Bug
**Affects Versions:** 0.4.0
**Reporter:** Gautam Borad
**Assignee:** Gautam Borad
**Fix For:** 0.5.0
**Attachments:** RANGER-284-Escape-HTML-before-displaying-to-prevent-.patch

*Steps to reproduce*
- Set user agent to something like this: "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0) <script>alert(1);</script>"
- Try to login to policy admin with an incorrect username/password
- Now login as admin user
- Go to Audit tab --> Login Sessions
- You will notice the failed logins displayed
- Click on the failed login session id
- Click Login sessions
- You will notice a Javascript popup alert (entered in the user agent)
- Expected Result*
- Unauthorized users should not be able to change the behavior of the application
- Actual Result*
- Unauthorized users are able to put javascript code that can be executed in admin users context
- Fix*
- Sanitize the user input data and any data comes from user.
DISTRIBUTIONS SPECIFICS
not in every environment
Distribution specifics

- Distros specifics
- User ifaces
- Hadoop
- Admin ifaces
- external ifaces
Basic distinction

cloud based

hosted locally
Distros

How long does it take to create a new distro version?

How many components are outdated at that time?

How long does it take to deploy a new distro at a company?

How many components are outdated at that time?

Most cases:
• MAJOR – ca. 1 year
• MINOR – ca. 3 months
• PATCH – ca. 1-2 months (differs much)
Hortonworks HDP components by version

http://hortonworks.com/hdp/whats-new/
Distros

Old components with known issues

• Old OS components (java, php, ruby, etc.)
• Old OS components (e.g. old tomcat used by Oozie and HDFS)
• Old Hadoop components (e.g. old Hue, Ambari, Ranger)

Default passwords

Default configuration
Vulnerability timeline

1. vuln found (e.g. Ambari)
2. Hadoop patched
3. distro update
4. deployment

Full disclosure? → Responsible Disclosure?

1. vuln found (e.g. jQuery)
2. jQuery patched
3. Django patched
4. Hue update
5. distro update
6. deployment

Full disclosure? → Responsible disclosure?
Distros

Old components with known issues

Default passwords

• SSH keys configured but default passwords still work
• Default mysql passwords, NO mysql passwords

Default configuration
Old components with known issues

Default passwords

Default configuration

- No network level hardening
- No HTTP hardening (clickjacking, session mgmt, errors)
- Hue uses Django with DEBUG turned on by default
- „Hacking virtual appliances” by Jeremy Brown
Default configurations sucks

**X-Frame-Options:** ALLOWALL

You're seeing this error because you have `DEBUG = True` in your Django settings file. Change that to `False`, and Django will display a standard 500 page.
EXTERNAL INTERFACES

For clients or whatsoever
External interfaces

- Distros specifics
- Hadoop
- Admin ifaces
- User ifaces
- external ifaces
External

- More than 25 internal Apache apps/modules
- Vendor/distro specific apps/interfaces
- Popular monitoring: Ganglia, Splunk
- Auth providers: LDAP, Kerberos, OAuth
- Many apps, many targets
SUMMARY

ways to protect your big data environment
Ways to protect your Hadoop environment

- Excessive network access
  - Keep it super tight!
- Excessive user permissions
- Typical web vulnerabilities
- Obsolete software
- Distros dependent vulnerabilities
- External system connections
Ways to protect your Hadoop environment

- Excessive network access
- Excessive user permissions
  - Map business roles to permissions
- Typical web vulnerabilities
- Obsolete software
- Distros dependent vulnerabilities
- External system connections
Ways to protect your Hadoop environment

- Excessive network access
- Excessive user permissions
- Typical web vulnerabilities
  - Pentest it! Introduce application independent security countermeasures
- Obsolete software
- Distros dependent vulnerabilities
- External system connections
Ways to protect your Hadoop environment

- Excessive network access
- Excessive user permissions
- Typical web vulnerabilities
- Obsolete software
  - Make a list of all components. Monitor bugtracks and CVEs.
- Distros dependent vulnerabilities
- External system connections
Ways to protect your Hadoop environment

- Excessive network access
- Excessive user permissions
- Typical web vulnerabilities
- Obsolete software
- Distros dependent vulnerabilities
  - A pentest after integration is a must. Demand security from software suppliers.
- External system connections
Ways to protect your Hadoop environment

- Excessive network access
- Excessive user permissions
- Typical web vulnerabilities
- Obsolete software
- Distros dependent vulnerabilities
- External system connections

- Make a list of all external system connections. Do a threat modeling and pentest corresponding systems.
Thank you

Contact me for additional materials

jakub.kaluzny@securing.pl