Securing Big Data at Rest with Encryption for Hadoop, Cassandra and MongoDB on Red Hat.

Alex Gonzalez | Software Engineer
Content

• What is Big Data
• 3 important No-SQL players
• Who uses Big Data
• Use Cases
• Encryption Solutions and its demo
• Navigator Encrypt
• Performance
• MongoDB, Hadoop and Cassandra Encryption
• Conclusion
What is BigData?

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured.
3 important NoSQL players

- **Hadoop**: Is a framework that allows for the distributed processing of large data sets across clusters of computers.

- **Cassandra**: A database with high availability, linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure.

- **MongoDB**: A scalable and high-performance, high availability, and easy scalability open source database designed to handle document-oriented storage.
<table>
<thead>
<tr>
<th>FINRA</th>
<th>SFR</th>
<th>EQUIFAX</th>
<th>TELKOMSEL</th>
<th>MasterCard</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Walt Disney Company</td>
<td>CapitalOne</td>
<td>BT</td>
<td>TATA Sky</td>
<td>comparethemarket.com</td>
</tr>
<tr>
<td></td>
<td>omneo</td>
<td>PCTEL®</td>
<td>maine stad.de</td>
<td>CONCUR</td>
</tr>
</tbody>
</table>

© Cloudera, Inc. All rights reserved.
<table>
<thead>
<tr>
<th>Company</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetLife</td>
<td>Building a single view of 100M+ customers across 70 systems in just 90 days.</td>
</tr>
<tr>
<td>Otto</td>
<td>Offering one-to-one shopping for 30M shoppers across 2M products.</td>
</tr>
<tr>
<td>Bosch</td>
<td>Creating new businesses by connecting sensors with real-time analytics.</td>
</tr>
<tr>
<td>City of Chicago</td>
<td>Delivering a unified view of city operations on a real-time geospatial platform</td>
</tr>
<tr>
<td>Expedia</td>
<td>Making travel planning easy, fast and highly personalized for millions of customers</td>
</tr>
<tr>
<td>Crittercism</td>
<td>Improving mobile app quality, 4 billion times a day</td>
</tr>
<tr>
<td>Forbes</td>
<td>Delivering a custom CMS in 2 months, and a new mobile site in 1 month.</td>
</tr>
<tr>
<td>ADP</td>
<td>Keeping 41,000 clients happy with its mobile app, personalized for over 1 million users</td>
</tr>
</tbody>
</table>
Application areas

• Business Intelligence, Analytics & Performance Mgmt
• Advertising, Sales & Marketing
• Advertising Network or Exchange
• Monitoring and Security
• Social
• Education and Training
• Data and Document Management - Financial, Health, etc.
• Music
• Video
• Gaming
Open source encryption solutions

dm-crypt
A transparent disk encryption subsystem

eCryptfs
eCryptfs is a POSIX-compliant enterprise cryptographic stacked filesystem for Linux.

Both are supported at Ubuntu, SLES, RedHat, Debian and CentOS.

Red Hat 7.x and CentOS 7.x are not supporting ecryptfs anymore.
eCryptfs and dmcrypt-demo
Cloudera Navigator Encrypt

Provides massively scalable, high-performance encryption for sensitive data. It leverages industry-standard AES-256 encryption and provides a transparent layer between the application and filesystem.
Navigator Encrypt Performance

Performance cost is ~5% to ~10%

mongodB

{ nThreads: 32, fileSizeMB: 1000, r: true }

new thread, total running : 1
Not-encrypted: 2380 ops/sec 9 MB/sec   Encrypted: 2479 ops/sec 9 MB/sec
Performance cost: 4.15%

new thread, total running : 2
Not-encrypted: 3011 ops/sec 11 MB/sec   Encrypted: 3160 ops/sec 11 MB/sec
Performance cost: 4.94%
MongoDB Demo
Navigator Encrypt Profiles

DMCRYPT

Navigator Encrypt works differently when creating ACLs for Java processes because the binary executed is the Java executable and Java can receive different jars.

In that case, you need to specify a profile, which contains all the options that Java receives when it gets executed. Using that profile, you can set which java application will access the data.
Hadoop Encryption

Navigator Encrypt Profiling - Obtaining the PID

```
[root@hdfs-2 ~]# ps aux | grep datanode
```

```
hdfs   7910  0.5  3.3 1649284 257040 ?   Sl   11:41   0:25 /usr/java/jdk1.7.0_67-cloudera/bin/java -Dproc_datanode -Xmx1000m -Dhdfs.audit.
logger=INFO,RFAAUDIT -Dsecurity.audit.logger=INFO,RFAS -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/var/log/hadoop-hdfs -Dhadoop.log.
file=hadoop-cmf-HDFS-1-DATANODE-hdfs-2.vpc.cloudera.com.log.out -Dhadoop.home.dir=/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/lib/hadoop -
-Dhadoop.id.str=hdfs -Dhadoop.root.logger=INFO,RFA -Djava.library.path=/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/lib/hadoop/lib/native -
-Dhadoop.policy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -server -Xms1073741824 -Xmx1073741824 -XX:+UseParNewGC -XX:
+UseConcMarkSweepGC -XX:-CMSConcurrentMTEnabled -XX:CMSInitiatingOccupancyFraction=70 -XX:+CMSParallelRemarkEnabled -XX:
OnOutOfMemoryError=/usr/lib64/cmf/service/common/killparent.sh -Dhadoop.security.logger=INFO,RFAS org.apache.hadoop.hdfs.server.datanode.DataNode
```
Hadoop Encryption

Navigator Encrypt Profiling

```
[root@hdfs-2 ~]# navencrypt-profile -p 7910
{
  "uid":"496",
  "comm":"java",
  "cmdline":"
}

[root@hdfs-2 ~]# navencrypt-profile -p 7910 > profile.txt
```
Hadoop Encryption

Adding a Navigator Encrypt ACL

[root@hdfs-2 ~]# navencrypt acl --add --rule="ALLOW @hdfs * /usr/java/jdk1.7.0_67-cloudera/bin/java" --profile=profile.txt

Type MASTER passphrase:
1 rule(s) were added
Hadoop Encryption

Verify Navigator Encrypt ACL Addition

[root@hdfs-2 ~]# navencrypt acl --list --all

Type MASTER passphrase:

#  Type  Category  Path  Profile  Process
1  ALLOW  @hdfs  *  YES  /usr/java/jdk1.7.0_67-cloudera/bin/java

PROFILE:
{
Hadoop Encryption

Stopping the cluster
Hadoop Encryption

Stopping the cluster
Hadoop Encryption

Navigator Encrypt Data Encryption

root@hdfs-2 ~]# navencrypt-move encrypt @hdfs /data/dfs/dn/current/ /mnt/mountpoint/

Type MASTER passphrase:

Size to encrypt: 12 KB
Moving from: '/data/dfs/dn/current'
Moving to: '/mnt/mountpoint/hdfs/data/dfs/dn/current'

100% [=================================================================>] [ 345  B]

Done.
Hadoop Encryption

Starting the Cluster

![Cluster Management Screen]

- **Cluster 1** (CDH 5.4.0, Parcels)
  - **1** Hosts
  - **2** HDFS-1

### Cloudera Management Server
- Cloudera Manager

### HDFS-1 Actions
- Start
- Stop
- Restart
- Rolling Restart
- Rename
- Delete

---

![Command Details: Start]

- **Command Details: Start**
  - **Command**: Start HDFS-1
  - **Context**: Finished
  - **Status**: Started at May 20, 2015 1:09:33 PM PDT
  - **Ended at**: May 20, 2015 1:09:57 PM PDT
  - **Details**: Successfully started HDFS service

- **Command Progress**: Completed 1 of 1 steps
- **Child Commands**
Hadoop Encryption

HDFS Test

[root@hdfs-2 ~]# su - hdfs
[hdfs@hdfs-2 ~]$ touch file.txt
[hdfs@hdfs-2 ~]$ hdfs dfs -mkdir /data/
[hdfs@hdfs-2 ~]$ hdfs dfs -copyFromLocal file.txt /data/file.txt
[hdfs@hdfs-2 ~]$ hdfs dfs -ls /data/

Found 1 items
-rw-r--r--   2 hdfs supergroup 0 2015-05-20 13:50 /data/file.txt
Cassandra Encryption

# ps aux | grep cassandra
root  15109  22.4  27.0  6347932  4143708 pts/0  SLl   00:22   0:08 java -ea -javaagent: /apache-......

# navencrypt-profile --pid=15109 > cassandra.profile

# navencrypt acl --add --rule="ALLOW @cassandra * /usr/lib/jvm/java-6-oracle/jre/bin/java" --profile=cassandra.profile

# navencrypt-move encrypt @cassandra /var/lib/cassandra/ /mnt/encrypted-mountpoint
Thank you
alex.gonzalez@cloudera.com