School of Informatics & IT
TEMASEK POLYTECHNIC

Data Security Fundamentals
CBG1C02
Assignment

Submitted by

MEENAKSHI DEKSHINAMURTHY
(2080548G)

30 Aug 2020

Temasek Polytechnic
School of Informatics & IT Specialist
Diploma in Big Data Management
AY2020/2021 Apr Semester
Data Security Fundamentals
CBG1C02
Assignment

Intake: 11

Submitted by:	Meenakshi Dekshinamurthy / 2080548G
Date:	30 / Aug / 2020

“By submitting this work, I am declaring that I am the originator(s) of this work and that all other original sources used in this work has been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.”

Signature of Student: Meenakshi D

Table of Contents

PART -1

1. INSIGHTS AND REFLECTION

In today's Information Age, data is priceless. Especially for, a company whose business is built around data, the data is rightly like gold. Unquestionably, these companies that handle personal and sensitive data must be fully aware of the threats and vulnerabilities that could compromise data.

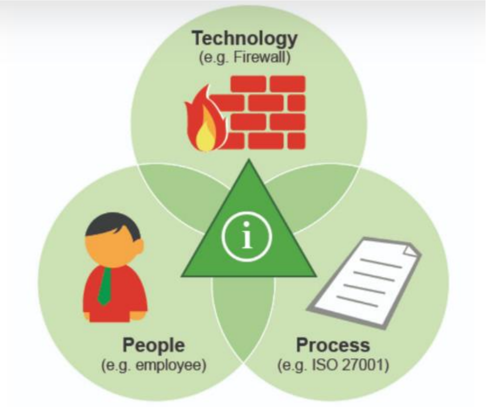
1.1 Thoughts about the incident

In the event of security incident, the detection, identification, mitigation and recovery or correction are of significance which I think BizOnline company has managed to carry out quite well. It is highly appreciable that the company had a forethought to employ a non-technical means of detecting the security threat by just watching out in the online forums. Also, the company's incident response is very professional and post-haste. Any further delay could have had serious repercussions to the company's reputation and revenue. The PR and security head working together is the very right thing to do. The whole investigation process as it unfolds is very systematic and well planned. The calculated step to first identify if the threat or loss is real, is a wise move. The hacker was also handled rationally rather than emotionally.

It is understandable that during an acquisition, there could be minor oversights or blind spots and there is always room for improvement. It might have been better if the company had taken time to understand the various network systems and security systems in place in the acquired company. Since there were taking ownership of the company , it would have been better if the company had invested resource on evaluating and training the security/IT staff to make them understand the importance of the security of data and how it is always possible to get impacted by a security incident. It's probably the lack of awareness of the acquired company that made them just remove the DNS entry and leave the servers without any maintenance. The company's move to strengthen the security and refine the procedures and exert a tighter hold on the acquired company, shows that it's on the right path for future.

1.2 Are there some insights you gained from the incident?

With more and more young intellects involvement in hacking activities, it is imperative that it is almost impossible to avoid the occurrence of a security threat altogether. Even a slight non-compliance of policy, procedure or a negligence in the company could lure the hackers. The risk is higher especially in the case of mergers and acquisitions. The companies should take all possible steps to identify such risks and assess their impact and implement appropriate preventive controls. The acquired company's standards, procedures, infrastructure and network, third party vendors and policies, third party applications and systems, if any, have to fully studied. Apart from the technical security controls like firewall, antimalware, scanning and monitoring, it is highly recommended to implement strict process adherence procedures and adoption of standard security policies and awareness training as a means of preventive control. This incident shows that there are three components people, process and technology that are important for a company's security. Any technology or process designed is meaningless if the employee is not aware and committed.



Source: https://www.securemetric.com/article/2013/preventive-security-vs-detective-security/

Even Data at rest has to be protected by encryption or masking. This reduces the impact of a security event by making the sensitive data unavailable for rogue uses and safeguards the privacy and interests of the owner.

If the acquired company is in a different country, the legal procedures and statutory requirements must be clear.

1.3 What do you think about the consequences?

Security Incidents leads to Emotional stress and Financial stress. Loss of reputation which is irreparable and loss of revenue as well. Loss of confidence in the company by the customer is another major consequence which for a growing company could be detrimental. The Company may lose current and potential customers. The security incident response costs lot of invaluable man hours and investigation expense. The employee in every connected department will have his routine operations disrupted by the investigation. Further the stress of legal proceedings and the compensations like regulatory fines, if any, will cost the business heavily. The cost of investment in security has to be increased because the company cannot afford to expose itself to yet another security incident. It will be hard pressed to invest in right security controls and security professionals to safeguard the company's interests and the client's interest.

2. HYPOTHETICAL QUESTION

If you had $1 million to spend on security controls, which of the controls (preventive controls, detective controls or corrective controls) would you spend most and least

As per the old saying, I would stick to "Prevention is better than cure" but security controls are essential to mitigate and reduce the risk to a company's asset as well. As a general rule the security controls designed that prevent breaches are cost effective than ones that detect or correct the breach.

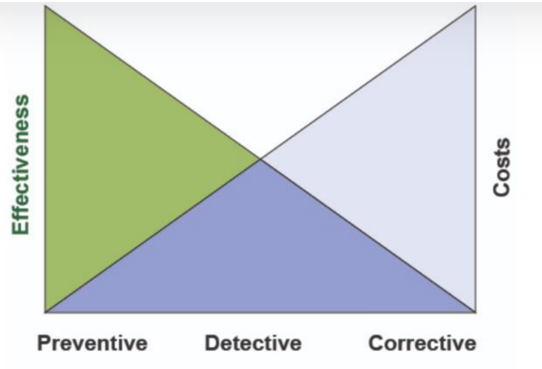


Figure 1: Cost & effectiveness against types of control (Source: Noticebored Ltd)

Source:https://www.securemetric.com/article/2013/preventive-security-vs-detective-security/

The hackers are overriding the security professionals. Before a solution is found for existing security threats, new threats are imminent for the online business. So, it sounds like it is more important to have a strong corrective or recovery control more than anything else.

The organization should do risk assessment and depending on the risk tolerance for confidentiality, integrity, or availability for a business, one or more combination of the controls is recommended. In the case of incident which involves stolen data, the data lost is almost irrecoverable. Unless the security incident is detected immediately, it is going to be almost impossible to ensure that data is not compromised. There is no corrective control available. Lost data is loss of reputation or business opportunity. A data at rest, the data stored on premise and HDFS should be protected by masking or encryption. That could render the hacker's effort useless even if he manages to gain access to it. If the data integrity is compromised, data backup, which is a corrective control would help to mitigate loss. Data in transit also can be protected by wired encryption. So BizOnline needs a strong preventive security control and in case an incident occurs a robust detective control like auditing and regular scanning.

I would opt to invest in adaptive Artificial Intelligence based security controls that could be implemented using Machine Learning Apps developed in Apache mahout. The cost of implementation of preventive control and detective control cannot be greater than the loss of reputation and even loss of business.

For the BizOnline, I would recommend more investment on prevention and detection control and to some extent on corrective control.

PART -2

3. IMPLEMENTING ACCESS CONTROL AROUND THE HADOOP CLUSTER

To list the possible users/groups and the access needed, we need to understand the data flow through Hadoop cluster.

Hadoop architecture has the staging DB area and production DB area. The staging area is where the data is loaded or ingested into HDFS without any extraction or modification. There will be continuous updates to it depending on the source of data like social media comments or web server logs using tools like Apache Kafka or other types of structured and unstructured data that need to be loaded to HDFS. So the users from the client side, access various services and write this data into HDFS. This data is cleaned and transformed by IT specialist and loaded into the production DB area for the data consumers like data scientists, analysts and long time storage**.**

HDFS and its users:

The OS user who starts the Namenode will be the superuser of HDFS. The hdfs user runs HDFS: NameNode, DataNodes, Standby NameNode, yarn runs YARN: ResourceManager, NodeManager, Mapred runs YARN: MapReduce, JobHistory Server. HDFS implements POSIX file permissions Read (4), Write (2), Execute (1) and ACLs to modify the permissions. The file created has default permission of 644 and directory has 755. Depending upon the role and responsibilities, the users are grouped.

Example: All the PR department users can be under a PR_group.

Sales under sales_group.

The Compliance Team under compliance_group.

The IT Specialists under it_group.

The Compliance may need read access to all the data to check the security and privacy compliance. The IT specialist group needs read and modify access to transform the data. The PR department may not need access to data. Sales department may need only read access.

Other Hadoop Components and its users:

Each component of the Hadoop ecosystem will have users and administrative users. The administrative users of the component manage the component's users and roles.
Example: Users of Hive will have roles like public and admin. Only the users with admin role can run administrative commands. Each user will be the owner of the table created by him and will have control over it (Discretionary Access Control). He can create, update, delete, view, and insert in his table and he can grant access to others. A table in HDFS is directory with zero or more files.

Roles are nothing but privileges. These privileges are assigned to users

Further the Database users have to be grouped to staging and production DB users and created groups and accounts to access the staging area DB and production DB in HDFS.
Typical Scenario: When a client user authenticated by the Client OS writes data in HDFS, the user is the owner of the file it created. And the group is the client users OS group. Say, (staging_user1, staging_group1 – the users belonging to a particular domain of client who ingest their data into HDFS.) The staging_user1 who ingests the file will be the owner of the file, the staging_group1 is the group where staging_user1 also belongs. The IT specialist may be in the staging_group1 and do the data transformation. He will have access to pig or Hive. The transformed data after approval by database administrator can be loaded into production DB. The IT specialist will have create and modify access to production db.

Production_user1, production_group1 (the users belonging to the same client domain who are the consumers of the transformed data). They can have read access to production area data but cannot change it. Any change to the production area needs proper approval process.

As mentioned previously, a complete security solution to Hadoop needs

Authentication: Authentication is the process of identification and verification of the user. The user identified by his username, is really who he claims himself to be is verified with his password. The default authentication is the “simple authentication” which depends on the client OS. This easily allows any user to impersonate users in hdfs.

Authorization: Authorization is the process of controlling the level of access of each user by restricting the resources and services exposed to him. HDFS provides directory and file level authorizations through file permissions. All the data is stored in files. So, the user has access to whole file, or he does not. These permissions can be controlled using ACLs by using the permission bits. There is no granular access in default HDFS. Authorization is an effective preventive and mitigative control since unauthorized users cannot access the data , even if they manage to impersonate a user.

Auditing: Hadoop audit logging can be enabled to log all access activity in HDFS, Mapreduce job logs, Hadoop daemon logs and they can be analyzed as a means of detection and prevention of breaches. Even authorized users can perform some unauthorized activities. So, auditing logs have to be monitored.

Encryption/masking as part of preventive control: Data at rest can be protected by masking or encryption. Data in transit also can be protected by wire encryption. The inter node and the inter process communication need to be encrypted. The read write between client and datanodes needs encryption. HDFS can be used with Hadoop's Key Management Service (KMS), or integrated with third party key management services for encryption.

4. SECURITY RISK AND PROPOSED CONTROLS

4.1 Security Risks

Hadoop was designed keeping in mind, the handling of huge public data in a trusted user environment. Hadoop by itself does not have any security feature. The eco system components of Hadoop like Kafka, Splunk, Pig, Hive, Hbase, Mahout all support various functions but have no security control. Components like oozie submit jobs on behalf of the users. In this scenario with varied users accessing data in HDFS from various nodes and applications, makes it difficult to control data access. The distributed computing and fragmented data make securing Hadoop a difficult task. By default, Hadoop does not have the ability to verify the identity of the client or the services running. It relies on the authentication of the client system. So, any user in client with the username same as that in HDFS will have access to that HDFS username account. If that impersonated user is of the superuser group, then it could be disastrous!

Also, such impersonated user can submit any job and pass it to the task manager for execution. It is not possible to safeguard the confidentiality of the data if user cannot be authenticated.

Both data and analytic process on premise and in the fly need to be secured.

Multi Layers of security needs to be provided to completely secure Hadoop.

4.2 Security Controls

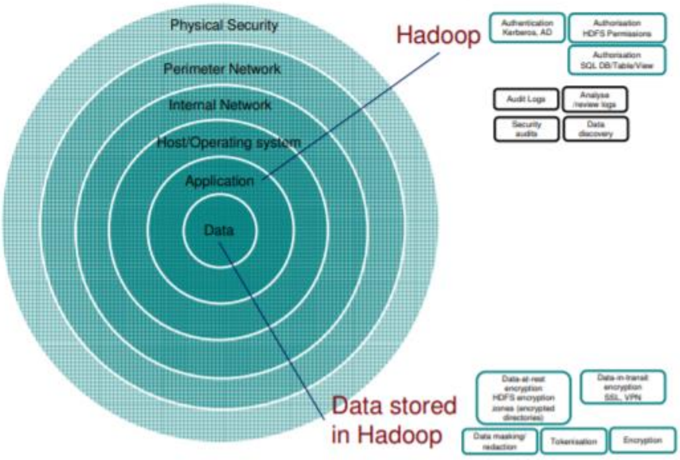
Kerberos Authentication:

Kerberos distributed authentication is to be implemented. Kerberos uses a ticket mechanism to implement authentication. It secures Hadoop by controlling how systems and users access Hadoop.

1. Establishing identity for clients, hosts, and services
2. Prevents the basic issue of impersonation and passwords are not sent over wire
3. Integrates with LDAP and Active Directory making it suitable for enterprise use
4. More details are available on data access and job execution allowing for granular auditing

Centralized Authorization and Auditing:

Instead of securing the various inter-node and inter-process communications separately at each Hadoop component, centralized control gives easy implementation and less load on OS resources. Apache Knox and Apache Ranger offer the centralized security control needed



Apache Knox is preferred as a gateway for perimeter security control and centralized authentication and is Kerberos compatible. At the entry point, it decides which user can access Hadoop cluster and protects the network details.

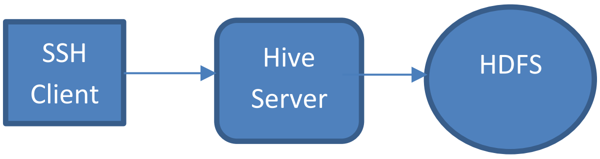
It limits the number of nodes which need direct access to Name node and hence reduces the chances of security breach. Knox has the capability to audit also. At the OS level, Host Intrusion Detection and at Network level, Network Intrusion detection systems can be implemented.

Apache Ranger to manages access control through a user interface that ensures consistent policy administration across Hadoop data access components. The policies are applied through a lightweight Ranger Java plugins, which run as part of the same process as the namenode (HDFS), Hive2Server (Hive), HBase server (Hbase) or any other service for a Hadoop component. It provides fine grain access control to Hadoop components which protects sensitive data. Using the Apache Ranger administration console, users can easily manage policies around accessing a resource (file, folder, database, table, column etc.) for a particular set of users and/or groups and enforce the policies within Hadoop. Ranger also provides dynamic data masking for data in transit.

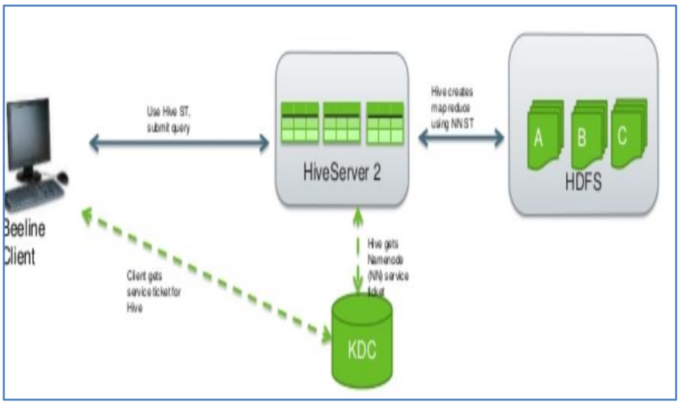
Source: https://www.slideshare.net/hortonworks/hdp-security-overview

Sample Default Hadoop Access

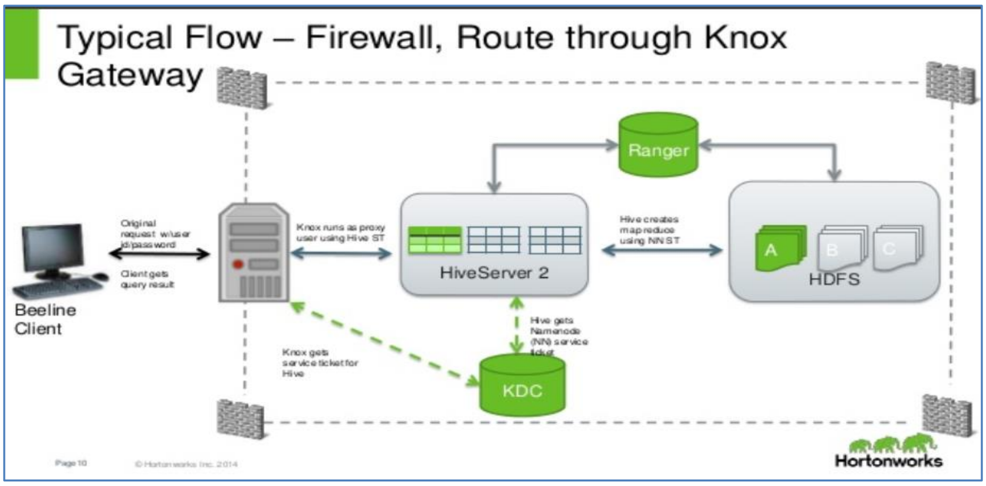
Without Authentication



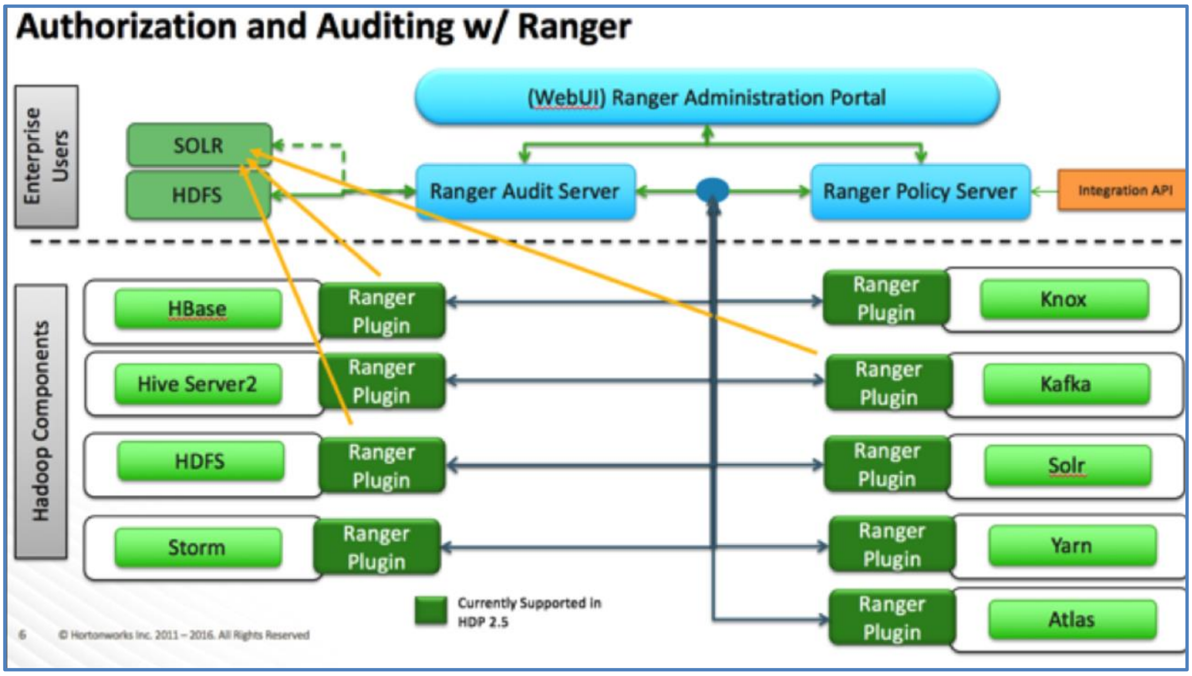
With Kerberos


Use Hive ST, submit query
HiveServer 2
Hive creates mapreduce using NN ST
A
B
C
HDFS
Beeline Client
Client gets service ticket for Hive
Hive gets Namenode (NN) service ticket
KDC

With Kerberos + Knox and Ranger


Typical Flow – Firewall, Route through Knox Gateway

Ranger
Original request w/user id/password
Client gets query result
Knox runs as proxy user using Hive ST
HiveServer 2
Hive creates mapreduce using NN ST
A
B
C
HDFS
Beeline Client
Knox gets service ticket for Hive
Hive gets Namenode (NN) service ticket
KDC
Page 10
© Hortonworks Inc. 2014
Hortonworks



Authorization and Auditing w/ Ranger

6 © Hortonworks Inc. 2011 – 2016. All Rights Reserved

5. REFERENCES

1. https://www.wisdomjobs.com/e-university/hadoop-tutorial-484/hadoop-security-14836.html
2. https://learning.oreilly.com/library/view/professional-hadoop/9781119267171/c06.xhtml
3. https://www.zdnet.com/article/cyber-security-spending-how-to-know-when-enough-is-enough/
4. https://www.techrepublic.com/article/11-security-questions-to-consider-during-an-it-risk-assessment/
5. https://learning.oreilly.com/library/view/information-security-a/9781849287418/xhtml/halftitle.html
6. https://www.techrepublic.com/article/nist-cybersecurity-framework-the-smart-persons-guide/
7. https://www.techrepublic.com/article/data-breaches-may-cost-less-than-the-security-to-prevent-them/
8. https://learning.oreilly.com/library/view/business-models-for/9781491947036/ch01.html
9. https://www.forbes.com/sites/allbusiness/2018/11/11/data-privacy-cybersecurity-mergers-and-acquisitions/#7a92f8c672ba
10. https://learning.oreilly.com/library/view/hands-on-incident-response/9781780174204/13_chapter01.xhtml#h16
11. https://learning.oreilly.com/library/view/threat-modeling-designing/9781118810057/9781118810057b02.xhtml