# Integrating Security into the Big Data Ecosystem

Anne M. Tall
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, FL, USA
anne.tall@knights.ucf.edu

Cliff C. Zou
Department of Computer Science
University of Central Florida
Orlando, FL, USA
changchun.zou@ucf.edu

Jun Wang
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, FL, USA
jun.wang@ucf.edu

***Abstract*— This paper provides an overview of the security service controls that are applied in a big data processing (BDP) system to defend against cyber security attacks. We validate this approach by modeling attacks and effectiveness of security service controls in a sequence of states and transitions. This Finite State Machine (FSM) approach uses the probable effectiveness of security service controls, as defined in the National Institute of Standards and Technology (NIST) Risk Management Framework (RMF). The attacks used in the model are defined in the ATT&CK® framework. Five different BDP security architecture configurations are considered, spanning from a low-cost default BDP configuration to a more expensive, industry supported layered security architecture. The analysis demonstrates the importance of a multi-layer approach to implementing security in BDP systems. With increasing interest in using BDP systems to analyze sensitive data sets, it is important to understand and justify BDP security architecture configurations with their significant costs. The output of the model demonstrates that over the run time, larger investment in security service controls results in significantly more uptime. There was a significant increase in uptime with a linear increase in security service control investment. We believe that this result supports our recommended BDP security architecture. That is, a layered architecture with security service controls integrated into the user interface, boundary, central management of security policies, and applications that incorporate privacy preserving programs. These results enable operationalizing BDP for sensitive data accessed in a multi-tenant environment.**

***Keywords—cybersecurity, big data, Hadoop***

## I. Introduction

Securing a Big Data Processing (BDP) system, whether in a cloud or a private, on-premise system is complex and different from other, computer, store and communication systems. The purpose of this paper is to detail unique security challenges and the security architecture strategy that must be taken. This is necessary to make the BDP large data lake of high volume, velocity, variety, veracity, and value (5-V) data available to a broad user group. These BDP security issues are expanding as more systems include data at various sensitivity levels and users with different clearance, authorization levels.

As described in the DoD Data Strategy [1], more military organizations are recognizing the value associated with analyzing large data sets for various purposes. This includes processing data from sensor systems to determine test results and auditing network traffic to determine cybersecurity status, for a broad range of government and business purposes. Operationalizing the security services that protect these large data sets is paramount.

Traditional security frameworks and architectures, such as Defense-in-Depth [2], are still applicable, however, these principles are implemented in a new manner. A unique characteristic of big data processing environments is that the analytics and tools introduced to derive meaningful insights from data are dynamic, uniquely developed for specialized purposes, and often open-source. In a sensitive, classified locked-down data processing environment, this type of dynamic introduction of executable code is akin to leaving the system open to the malware. The traditional closed, controlled approach can result in a substantial data-lake investment that is accessible by very few data analysts.

In this paper, we propose an approach to integrating security control services into BDP architectures to avoid this situation. This is intended to break down barriers to data access and advanced analytics. Our proposed approach is based upon work in this area by standards group such as the National Institute of Standards and Technology (NIST) [3] [4], open-source projects, and industry initiatives [5].

We validate the proposed approach by modeling cybersecurity attacks, as defined in the ATT&CK® framework [6] [7], against the recommended security service controls defined in the NIST Risk Management Framework (RMF) [8] [9]. Based upon probabilities that attacks will be successful and security services will appropriately defend the BDP system, our model generates a system degradation value, uptime, and conceptual costs associated for five configurations. This analysis can be used by the military, managers, engineers, and computer scientists to guide their big data system security strategy.

The unique contributions of our research, as documented in this paper, are to:

1. Propose and describe a layered BDP security architecture based upon a survey of approaches to Hadoop security,
2. Describe provisioning standard security service controls in this architecture,
3. Justify the proposed architecture by modelling the execution of cybersecurity attacks and use of security services to mitigate these attacks based upon five different cost models.

BDP systems are transitioning from large, monolithic High Performance Computing (HPC) systems to distributed clusters that execute in parallel. The fundamental design of BDP file management systems, such as the Hadoop Distributed File System (HDFS) is an architecture that allows for scale-out. This

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©2021 IEEE

is a design where the total storage capacity can dynamically expand by continually adding small, commodity servers as the data size grows. Prior HPC systems were designed around a scale up architecture, which involved adding more CPU cores, RAM and disk storage. Scaling up to ever more powerful computers is expensive. However, more end points presents more vulnerabilities both in the system and on the network.

The opportunities provided by open-source parallel processing BDP systems, such as Hadoop, are exciting and also complex due to the large number of components that comprise a BDP ecosystem, [10] [11]. Maintaining the security of these systems requires not only an understanding of the core storage, compute, and resource management components, but also an array of components that provide additional services such as high availability, management of high data volumes flowing into and out of the cluster, scheduling jobs, providing security and others. These applications provide different methods for accessing the same data. Therefore, it is critical that each component applies security in a consistent manner.

In a multi-tenant use case, stored data is shared across the organization (different mission/business groups and users) in a way that enables each organization to run their own applications (e.g., MapReduce programs, Pig jobs, Spark applications, HIVE, Hbase). Security services need to be configured so that each user is segregated from each other and able to access only their authorized data.

BDP systems, such as Hadoop, are designed for speed in storing and retrieving data, so the data is not normalized or indexed upon storage. This is quite different from traditional relational databases where a schema is imposed upon the data before it is stored. This schema often serves as the basis for security decisions. However, the concept of a schema is not present in the big data distributed file system paradigm. Security was not a priority in the original development of HDFS. Edge systems were added to provide the structure for queries against the BDP in a Structure Query Language (SQL)-like manner and the associated relational database model of security.

The results provide evidence of the value and critical nature of securing a BDP system prior to operationalizing it for processing mission critical sensitive data. As recently stated by LTGEN Berrier, Director Defense Intelligence Agency to the United States Senate in April 2021, “the United States will increasingly face advanced, persistent, and sophisticated malicious cyber activities emanating from a wide array of state and non-state actors.,” [12]. The need to protect BDP systems is now more pressing than ever.

## II. Background

Measuring the correct amount or level of cybersecurity that needs to be integrated into the architecture of large scale, diverse data processing systems has been a long-term challenge in the information security domain. Standards organizations have published guidance on cybersecurity measurement based upon best practice, consensus approaches, however many measurements remain subjective. Objective quantification given the dynamics associated with attacks and protection mechanisms continue to challenge computer, network system manager and administrators. The recently published IEEE Standard for Big Data Security [13] helps to improve the assessment of big data technology security protection mechanisms against business security. This standard defines a framework that consists of a portrait level and algorithm level approach. By standardizing business risk assessments, improvements can be made in sharing, evaluating, and predicting BDP risk posture and inheritance when interconnecting to other systems. However, the standard depends upon the assignment of risk based upon several subjective factors such as data sensitivity levels.

Other security measurement standards applied to the security architecture analysis included the Exploit Probability, Impact Factor, and Service Availability as defined by NIST [14] [15]. Quantifying these factors can be complex and selecting the correct scale based upon false precision can lead to inconclusive results, [16]. Therefore, in this analysis we apply the guidance to use simple metrics that help to quantify observations of attack and security service effectiveness. Although complex interrelationships of attack paths and redundant detection, correction systems exist, these more complex situations were not incorporated into the model. Therefore, the overall results of the model were used for broad recommendations for a layered approach to security, rather than requirements for specific security service control or mechanisms.

As more quantitative data on attack and security service control countermeasures is made available from test or real-world events, more complex interrelationships could be included in the model, such as the cyber-attack analysis conducted by Liu, Xing and Zhou, [17], using Continuous-Time Markov Chains (CTMC) to capture the interdependence of attacks. As more complex multi-step attacks are incorporated into a model, however the complexities associated with the details can limit the scope of the analysis which could limit the diversity of attacks considered and skew focus and acquisition towards a subset of the necessary security mechanisms. For example, Chen et al., [18], analyze vulnerabilities using a FSM approach and reached insightful conclusion on a few threat campaigns under analysis. Scaling out this of analysis at a high level of detailed fidelity while maintaining overall accuracy could be challenging.

## III. Approach To Security Service Integration

BDP system security is different from other data processing systems, e.g., relational databases. BDP processing is characterized as an ecosystem, in that the various components, such as the Hadoop software library and the accessories and tools provided by various Apache Software Foundation projects are independently developed capabilities, however they all work together to provide a complete data management and processing environment. This results in differences in the implementation, integration, and execution of security services in the following ways:

- Security services (mechanisms) need to be applied in a distributed manner in the data processing and compute, store layers, e.g., at the master node, each data node, and supporting ecosystem server (MapReduce, Spark, Hive).
- Security information, (policies and permissions), need to be managed centrally and distributed through trusted

methods to all the components in the big data ecosystem from the management layer.

- Security decisions, such as identification, authentication, access control and system and communication integrity, are made at all ecosystem components, not only by boundary, proxy servers at the gateway boundary layer.

Motivated by these differences and based upon our survey of BDP security research, an architecture that employs security services at layers within the big data ecosystem is recommend [19] [20] [21] [22]. This approach is depicted in Fig. 1. The diagram summarizes the layers of the BDP architecture, and the placement security services in each of these layers.

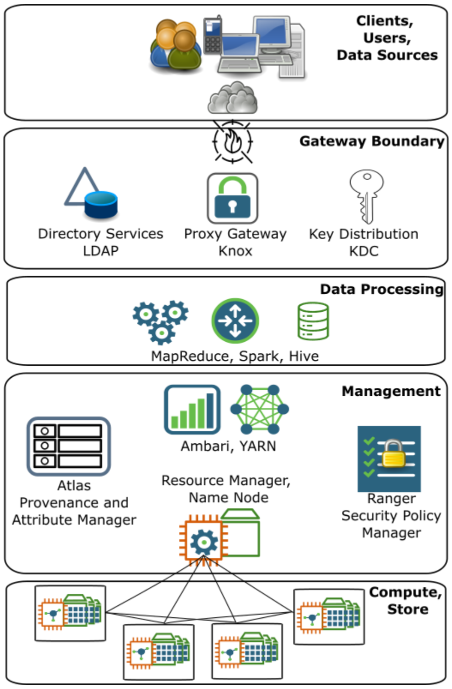

Fig. 1. BDP Security Architecture: recommended layered approach.

Not all of the NIST RMF security service control families are implemented by the security mechanisms in the Hadoop ecosystem. For example, Awareness and Training (AT), Physical and Environmental Protection (PE) and Personnel Security (PS) security service controls are mostly external to a BDP system and satisfied through procedures and policies. A map of the NIST RMF [8] security control families provided by the BDP security mechanisms in the architecture are shown in Fig. 2.

Researchers and open-source developers propose various components to secure the ingest, tracking, storing, analyzing, and producing summary reports with BDP systems. Integrated together, a layered, defense-in-depth solution is achieved. The following sections further describe the security services and mechanisms recommended for each architecture layer.

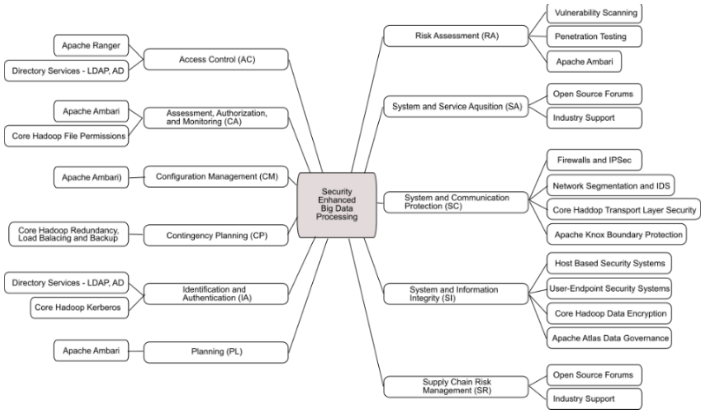

Fig. 2. Mind Map: security service control to BDP security mechanism map.

### *A. Gateway Boundary Layer*

The BDP gateway, boundary layer builds upon traditional network boundary protection by providing application-specific gateway proxy services. Also, identification, authentication, and access control services through either a BDP-specific or enterprise-integrated directory service is a critical part of the boundary security services.

Typical products at this layer include Microsoft Active Directory (AD) or Open Lightweight Directory Access Protocol (LDAP) with a Kerberos Key Distribution Center (KDC) and Apache Knox application gateway proxy server.

Kerberos is the authentication mechanism integrated and optionally configured in core Hadoop. The original Kerberos implementation was developed at MIT and is currently available as open-source software. A feature of Kerberos is that because it is based on symmetric-key, keys used to authenticate and encrypt connections are shared. Public Key Infrastructure (PKI)-based alternative approaches used in Transport Layer Security (TLS) use asymmetric keys managed by a Certificate Authority (CA) which overcomes the potential security challenge of shared keys; however, the processing level can be more intensive. Use of PKI to secure Hadoop has been investigated by researchers, [23], however, we would expect that the most implementations are using the default Kerberos services.

At this architecture layer Apache Knox provides a stateless reverse proxy for a single point of access to the Hadoop cluster. It provides authentication, auditing, authorization for external users. It reduces the number of access points and can provide a single URL for accessing Hadoop services. This can provide security by concealing of Hadoop cluster installation details and data. Knox works with the AD or LDAP server to authenticate users external to the perimeter [24] [25].

### *B. Data Processing Layer*

The data processing layer can consist of a wide variety of parallel processing and SQL to HDFS interfaces. Each of these can have their own Application Programming Interface (API) to authenticate and negotiate access to the distributed file system. This API authentication includes programs that stream data from external sources in to the BDP system.

Security at this layer depends upon configuration of access control, identification, authentication of permissions for users and their associated applications as well as the files (data and executables) permission settings. Researchers have proposed strategies to add security features to a library calls or modify to the data analysis programs (e.g., SQL on-Hadoop, MapReduce). A challenge with this strategy is that it needs to be coupled with strong controls that prevent the introduction of any unmodified or unauthorized analysis programs. Several important proposed concepts include query modification to extend access controls [26], rewriting queries to enforce privacy aware access controls [27], and splitting execution of MapReduce programs between private and public clouds based upon data privacy policies, [28].

Other techniques used to provide security at the data processing level include privacy preserving programs executing in parallel to mask sensitive data. Scaling out data anonymization techniques and tracking this as a sensitivity attribute enables enforcement of security policies so that sanitized data can be made available to users with lower authorization levels.

### *C. Management Layer*

A robust management layer depends not only upon core Hadoop services, but also key ecosystem products to provide configuration, security policy, provenance and attribute management. Many of the security service control families are achieved at this layer through robust management tools, such as open-source Apache Ambari and commercially supported systems, such as the Cloudera Manager. Researchers have reported on the performance gains in an optimized, configured system, [29].

In addition to management tools, the critical BDP management components include security policy management, such as with the Apache Ranger tool and data attribute life cycle, provenance management, such as with Apache Atlas.

Apache Ranger is the primary open-source framework for securing Hadoop. It manages the authorizations across the Hadoop ecosystem (HDFS files, Hive tables, etc.). Ranger uses Kerberos for authentication and TLS for encryption of data exchanged over the network. Highly granular, specific security policies can be defined and implemented across the ecosystem using Ranger [30] [31].

Data provenance is defined as the record of the source, processing, and overall lineage of the data. These metadata attributes that track data provenance are critical to big data systems. Traditionally data provenance is associated with audit logs and debugging. In Apache Atlas, data provenance can be expressed using a data model, business vocabulary, or other directed acyclic graphic terms. Making big data sets available for analytics in a secure manner requires tracking when process are executed that reduce the data sensitivity then updating the data provenance attribute in a trustworthy manner. Research has been published that describes using metadata tags to track processing provenance in this manner, (e.g., sanitization history) [32] [33].

### *D. Compute, Store Layer*

The current Hadoop architecture uses metadata to handle the distribution and load balancing blocks across the data nodes. Like other file systems, the Hadoop File System (HDFS) uses a POSIX style Access Control (AC). The Apache Ranger project provides the hooks, using software plugin programs that are installed on each component, to manage access on each node, including Name, Resource, Job History and Data Nodes. Therefore, security AC checking is extended into the core Hadoop system, in a consistent, centrally managed manner. This achieves layered defense-in-depth. Several projects and commercial tools leverage the Ranger hooks to facilitate metadata management e.g., Apache Atlas, UC Berkeley Ground, and Cloudera Navigator. This provides the opportunity to integrate additional metadata into the AC decision.

HDFS file system security is distributed across all the nodes in the Hadoop cluster. File permission settings are optional and by default disabled. When the file system permissions are disabled, anyone with access to the computer system node can do anything to the HDFS files. Also, encryption of files stored in HDFS is optional and disabled by default. Anyone with access to the local disk can read the unencrypted files. Data in the Hadoop cluster is exchanged in the clear, that is all network traffic is unencrypted by default.

The ability to disable file permissions, data encryption at rest and when exchange (transmitted over the network) between nodes highlights that security was added to Hadoop after initial development. Designing in security services at the beginning of the development process generally leads to an overall more secure design and reduces opportunities to bypass intentionally configured security services.

Basic security features are configured through settings in Hadoop XML files. Without these configurations, any HDFS account on any node in the Hadoop cluster is permitted access to their files anywhere in the cluster. Basic Hadoop operations where files are created in folders and Map-Reduce programs are executed with these files are open for any user to execute if the default security settings are not changed. The Hadoop fsck command allows users to know where blocks for any particular file are stored and can see the metadata to find all replicated copies of data. Cross system authentication is accepted, and users do not have to reauthenticate, e.g., provide a local system password, when reusing accounts across two systems in Hadoop.

## IV. Security Architecture Analysis

To analyze the proposed layered security service architecture, described in the previous section, a Finite State Machine (FSM) model of attacks and security service controls was developed. This model demonstrates the use of BDP systems security service controls to thwart the impact of cybersecurity attacks and increase uptime. Overall, it provides insights on the value of security service controls.

A linear increase in security mechanism investment and maintenance results in an exponential increase in uptime. The FSM used as the basis for the model is shown in Fig.3. It consists of 5 states and 10 transitions. In each state, the evaluation of different aspects associated with cybersecurity attacks and defensive security service controls was incorporated. Using best practices and reports from industry and academia, the likelihood of the attacks and defenses was considered. Specifically, random values, based upon the Binomial or Poisson distributions, were computed to represent the chance of attacks and likelihood the defenses were successful. The conditions and probabilities evaluated at each state are further described below.

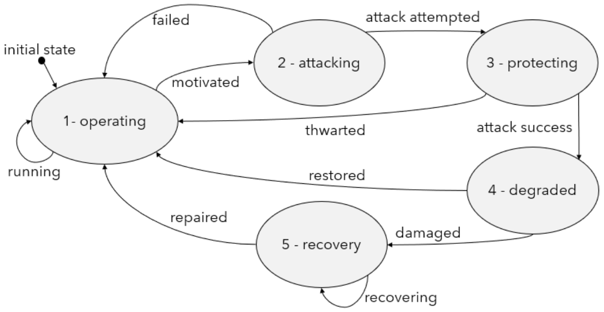

Fig. 3. Finite State Machine (FSM) Model: BDP cybersecurity attacks and protections.

### A. *Operating*

The initial state represents the BDP system in an active, operational state. The motivation of cybersecurity attackers are evaluated at this state. In this model, a transition out of the operating state was based upon two factors: (1) the value of the BDP system and data and (2) environmental conditions.

For the motivation based upon the value of the BDP system and the data it processes, we applied the Federal Information Processing Standards Publication 199 (FIPS PUB 199) standard System Categorizations (SC), [34]. The confidentiality, integrity, and availability (CIA) required of the system determine the SC, in accordance with the following formula:

SC information type = {(confidentiality impact), (integrity impact), (availability impact)}

The impact of loss of CIA is rated as:

- Low = limited adverse effect,
- Moderate = serious adverse effect, or
- High = severe or catastrophic effect.

For example, a system with classified military system performance data would have a system categorization of high in all three CIA areas (SC = high, high, high).

The other factor, environmental conditions, have been identified in industry and academic research reports as motivations or triggers for attackers. Environmental conditions that we identified during this research were:

- Domain Name Service (DNS) name - researchers have found that systems with a specific DNS name are specifically targeted by cybersecurity attackers, [35].

- Business Type - 90% of all attacks are about financial gain and espionage, so certain industries, such as government, healthcare and financial companies are more at risk, [36].

- Political Climate - the internal and external political climate surrounding the organization that owns the BDP, such as recent layoffs could increase the likelihood of an insider attack, [37].

- Media Attention - more media attention on the BDP owner, increases the likelihood of an attack, for example the COVID-19 pandemic has resulted in an increase in World Health Organization (WHO) related attacks, [38].

These areas were combined with the SC for an overall attacker motivation probability in the model. There are other factors that could be considered in evaluating the motivation of a cybersecurity attacker, however overall, most systems connected to the Internet today have experienced at least one or more cybersecurity attack, so we view an overall motivation factor of 80% or higher as reasonable, [39], and used for transition from state 1 to 2 (1-2) motivated.

### B. *Attacking*

This state represents the likelihood that a cyber security attack will be launched against, sent to, or executed on a BDP system. This state considers the likelihood and impact probability of approximately 350 different attack methods defined in the ATT&CK® framework.

The DoD Cybersecurity Table Top (CTT) Guidebook provides the definition of cybersecurity attack likelihood and impact that we used as the basis for assigning probabilities to the attacks, [40]. A scale of 1 to 5 was applied, with technically complex, low likelihood attacks were assigned probabilities in the 1 to 2 range and technically easy, well know and more highly likely attacks were assigned probabilities in the 4 to 5 range. A value of 3 was used for moderate complexity and likelihood. At this state, each attack was considered independent of the presence of security service controls or mechanisms that might thwart or otherwise neutralize and stop the attack.

The potential impact of the attack on the operation of the BDP system was also assigned on a scale from 1 to 5. A low value assigned for impact would indicate a cybersecurity attack would have little impact on the mission, for example, whereas a high value of impact could indicate a mission abort if the system were under cybersecurity attack. This value was used as the probability (p) in generating a Binomial distributed random value. If the resulting variate is "success" then the attack is considered "attempted" and there is a transition from state 2 to 3 (2-3) attack attempted to the protecting state. If the computed variate is "failed" then there is a transition back to the initial operating state.

### C. *Protecting*

For each attack represented as successful, we evaluated the corresponding security service control. The mapping between NIST RMF security service controls and cybersecurity attacks defined in ATT&CK® supported the assessment of protection measures, [9], [41].

This evaluation considered the likelihood a mechanism to provide the security service control was implemented and maintained. For a BDP system such as Hadoop, the likelihood of a particular mechanism would depend upon the maturity, support, and investment in securing the system to an operational business grade status. Five cost model configurations of a BDP system, based upon a Hadoop ecosystem, were defined and used in the model as summarized below:

- Cost Model A – Default Hadoop installation
- Cost Model B – Use of core Hadoop security services and Operating System (OS) security

- Cost Model C – Enhanced with open-source security systems
- Cost Model D – Industry supported security systems
- Cost Model E – Enhanced (e.g., secure cloud) industry managed services

The resiliency and completeness of the systems security increase from A to E, with E representing a complete, layered security architecture with managed security services.

Of the seven levels defined in the Common Criteria (CC) Evaluated Assurance Levels (EAL) five were used in the model as the basis for assessing the strength of the service and assigning probability values for likelihood of implementation, [42]. The CC EAL, as applied to the BDP system, is summarized listed below:

- EAL1: Functionally Tested
- EAL 2: Structurally Tested
- EAL3: Methodically Tested and Checked
- EAL 4: Methodically Designed, Tested and Reviewed
- EAL 5: Semi-formally Designed and Tested

The other consideration incorporated in the model for the probability of successful protection is the maintenance of the security service control. The DoD Cybersecurity Maturity Model Certification (CMMC), [43] and the related Carnegie Mellon Software Engineering Institute Capability Maturity Model (CMM), [44], were used to guide the assignment of probabilities to maintaining the security mechanisms.

Given the complexities of setting up an open-source Hadoop ecosystem with many options for injecting, transforming, processing and displaying big data, the balance between security and flexibility can easily be focused away from system and data protection.

Approaches as characterized by Cost Model A would represent less mature processes and practices that include design decisions that sacrifice security over flexibility. Whereas configurations represented by Cost Model D or E would exemplify more mature cybersecurity processes and procedures.

The average of the probability of implementation and maintenance was used as the probability input to compute a Binomial distributed variate. If the "success" of the control in preventing the attack is computed the state 3 to 1 (3-1) thwarted transition is executed and the state is transitioned to operating. If the security service control random value is computed as "unsuccessful," the attack is considered successful and there is a transition (3-4) to the degraded state.

### *D. Degraded*

The In the degraded state, the BDP system is considered compromised by the attack. The ability to recover or be resilient without incurring down time is evaluated in this state. Effectiveness of mitigations is evaluated. There may be impacts, such as performance degradations or defacements that could be considered as damaging the reputation of the BDP owner. The two conditions evaluated in the degraded state are: ability to operate degraded and the impact of the degradation.

The ability to operate degraded is the probability the BDP system can continue to operate in a configuration where the system has been subject to a cybersecurity attack, for example data is changed in an unauthorized manner, however, the processes continue to execute in a manner such that manipulated results can be detected and corrected. The ability to operate degraded is computed based upon the sophistication and impact of the attack and security service controls. Less technically sophisticated and low impact attacks with mature security controls that counter the attack increase the probability the system can operate degraded. The probability calculated based upon a combination of these factors is input to a Binomial distributed variate generator to determine the transition from the degraded state (4) to either the operating state (1) ("success") (4-1) or the recovery state (5) ("failure") (4-5). The formula used to determine if the system can operate degraded is:

*Operate Degraded Probability = ((1 - attack probability) +*
*(1 - attack impact probability) +*
*control implementation probability ) / 3* (1)

*Operate Degraded = True, when the Binomial distributed*
*random variable generated using the*
*Operate Degraded Probability is True (success)* (2)

The impact of the degradation is calculated based upon the degradation value assigned to each successful attack. The sum of the degradation value ranges from 0 to 20, based upon the maximum of the individual values (1 to 10) and an amplification value based upon the volume of attacks, [45]. Lower values are associated with attacks that result in minimal noticeable performance impacts and detected data damage or breach. Higher values are associated with more extensive performance slowdowns, data breaches and significant reputation damage, such as external web site defacements. The resulting degradation value is computed for each Cost Model A-E. The formulas (3) and (4) used to compute the degradation value in the model are:

*Degradation Value = Maximum (degradation value*
*assigned to each successful attack) + Amplification Value* (3)

*Amplification Value = Average Degradation Value for all*
*the successful attacks * Volume Score,*
*where the volume score ranges from 0 to 1*
*based upon the number of successful attacks* (4)

### *E. Recovery*

Like calculating the impact of the degradation, each NIST RMF security service control mapped to an attack has an average recovery time assigned. The amount of time associated with recovering roughly corresponds to the complexity of the attack and maturity of the security service. The unit of time used in the model is hours, with values for each successful attack ranging from a minimal amount of time (one hour) to 48 hours (2 days). The average recovery time from the reference spreadsheet is used as input to a Poisson distributed random variate computation. The resulting recovery time is then added to the total down time summed for each Cost Model. After recovery is complete there is a transition (from state 5 to 1) back to the initial operating state. The formulas (5) and (6) used to compute down time in the model are:

*Down Time = Maximum (Down Time assigned to*
*each successful attack) + Amplification Value* (5)

*Amplification Value = Average Down Time for all the successful attacks * Volume Score,*
*where the Volume Score ranges from 0 to 4*
*based upon the number of successful attacks* (6)

## V. Simulation Results

The result of running the model through 365 FSM cycles, 5 times, for each Cost Model A-E is shown in Table I and Fig. 4. The results illustrate that a linear investment in security mechanisms results in significant improvement in reducing down time. The first two models (A and B) which only use the default and basic Hadoop security services are not resilient to cybersecurity attacks and down the entire time. The mid-cost model (C) is down a significant amount of time and also operates in a degraded state. The models with the most robust security configurations (D and E) experience the least amount of downtime. However, the most robust configuration, also operates with the least amount of degradation. Clearly, for mission and business critical systems, where down time or operating in a degraded state can have a significant impact on readiness and business continuity, there is strong justification for the most robust security architecture configuration.

TABLE I. Degredation, Down Time , Up Time Averages

| Model | Cost Model ($K) | Degredation (Total / No./ Avg.) | Down Time (days) | Up Time (days) |
|---|---|---|---|---|
| A | 2,200 | 0 / 0 / 0 | 342 | 23 |
| B | 2,900 | 2 / 1 / 2 | 335 | 30 |
| C | 5,500 | 231 / 38 / 6 | 271 | 94 |
| D | 8,300 | 704 / 117 / 6 | 141 | 224 |
| E | 9.500 | 63 / 11 / 6 | 9 | 356 |

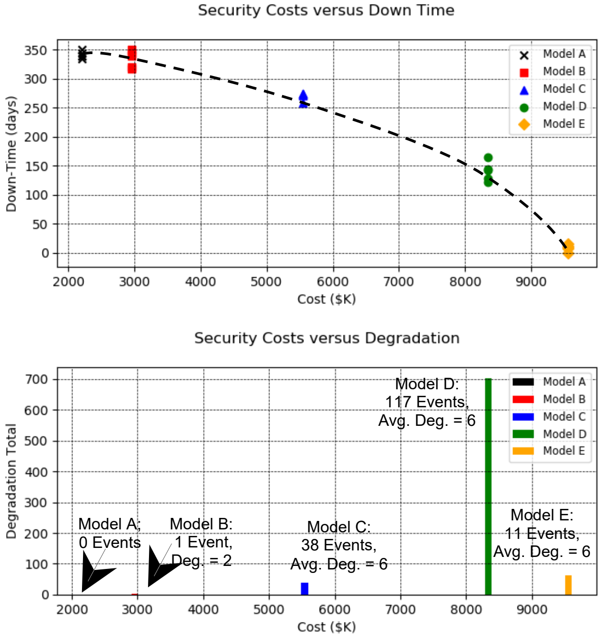

Fig. 4. Model Results: linear down time relationship and degradation impact total, number of degradation events and average degradation values noted.

Although the factors considered in this model are focused on the BDP system architecture, the concept of analyzing cybersecurity attacks and the resilience of the protection mechanisms to reduce down time could be applied to other information technology systems. This can help to support and justify investments in security mechanisms. Security system investment can be viewed as black hole, where an unlimited amount invested can appear to have negligible return on investment. However, this experiment demonstrated in an empirical manner that investment security mechanisms can have a significant increase in cybersecurity resiliency, i.e., resistance to cybersecurity attacks.

## VI. Summary Recommendations

In summary a layered security service control architecture is required for secure, multi-tenant BDP that includes:

- Enforcement of users and data sources authentication and access controls at the gateway (or proxy) boundary.
- Execution access control, identification, and authentication security services as well as control of process execution, such as incorporating privacy preserving programs as part of the data processing layer.
- Integrated security policy information and administration as part of the management layer, including data provenance and resource management.
- Policy decisions and enforcement on each compute-store node using local agent, client APIs synchronized with the security policy manager.

Complete layered approach of application of mechanisms that implement security service controls was demonstrated to have significant improvements in reducing the effectiveness of cybersecurity attacks based upon reasoned probabilities. In open source BDP systems, managers and administrators need to be proactive in configuring systems in a secure mode, since most default installations are unsecure, e.g., encrypting data at rest and over the network. Proactive identification and integration of open-source projects into the BDP ecosystem is required to complete the security architecture. The entire data processing lifecycle tool chain needs to be considered including implementing and tracking security policies using provenance tools and privacy preserving programs for example.

## VII. Conclusions

In conclusion this paper describes the research and analysis conducted on BDP security architectures. A model security architecture was presented along with a mapping to security service controls. It was demonstrated using a FSM model that linear investments in security mechanisms results in exponential improvements in reducing down time. Cybersecurity attacks are persistent and increasing, however BDP security mechanism implementation and maintenance can be complex and expensive when trying to integrate several independently developed open-source components. This analysis shows that to achieve an operational system capable of processing mission critical, for multiple users.

## Acknowledgment

This work was sponsored by the U.S. National Science Foundation (NSF) under Grants 1915780, 1717388, 1907765, and 2028481.

## REFERENCES

[1] Dept. of Defense, "DoD Data Strategy," Deputy Secretary of Defense, Washington, D.C., 2020.

[2] National Security Agency, Central Security Service, "Defense in Depth: A practical strategy for achieving Information Assurance in today's highly networked environments," Information Assurance Solutions Group - STE 6737, Fort Meade, MD, 2010.

[3] NIST Public Big Data Working Group, "Big Data Interoperability Framework," NIST, U.S. Dept. of Commerce, Gaithersburg, MD, 2019.

[4] NIST Big Data Public Working Group, "NIST Big Data Ineroperability Framework: Volume 4, Security and Privacy, Version 3, SP 1500-4r2," U.S. Dept. of Commerce, Gaitersburg, MD, 2019.

[5] Research Data Alliance (RDA), "Big Data Interest Group (IG)," [Online]. Available: https://www.rd-alliance.org/groups/big-data-analytics-ig.html.

[6] The MITRE Corporation, "ATT&CK®," The MITRE Corp., McLean, VA, 2021.

[7] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington and C. B. Thomas, "MITRE ATT&CK: Design and Philosophy," The MITRE Corp., McLean, VA, 2020.

[8] NIST, "Risk Management Framework," NIST, U.S. Dept. of Commerce, Gaithersburg, MD, 2021.

[9] MITRE-Engenuity, "Security Control Mappings: A Bridge to Threat-Informed Defense," The MITRE Corp., 15 Dec. 2020. [Online].

[10] J. Ronan and contributors, "The Hadoop Ecosystem Table," GitHub Pages, 2021. [Online]. Available: https://hadoopecosystemtable.github.io/.

[11] T. White, Hadoop: The Definitive Guide, 4th ed., Sebastopol, CA: O'Reilly Media, 2015.

[12] S. Berrier, "Statement for the Record: Worldwide Threat Assessment, Armed Services Committee, U.S. Senate," DIA, Washington, D.C., 2021.

[13] BDBSRAWG - Guide for Big Data Business Security Risk Assessment , "IEEE 2813-2020 - IEEE Standard for Big Data Business Security Risk Assessment," IEEE CTS/SC Standards Committee, New York, 2021.

[14] NIST Information Technology, "Measurements for Information Security," U.S. Dept. of Commerce, 15 Sep. 2020. [Online]. Available: https://www.nist.gov/cybersecurity/measurements-information-security.

[15] Y. Cheng, J. Deng, J. Li, S. DeLoach, A. Singhal and X. Ou, "Metrics of Security," NIST, Gaithersburg, MD, 2014.

[16] D. Flater, "Bad Security Metrics Part 2: Solutions," *IEEE IT Professional,* vol. 20, no. 2, pp. 76-79, 2018.

[17] Q. Liu, L. Xing and C. Zhou, "Probabilistic modeling and anlaysis of sequential cyber-attacks," *Wiley Engineering Reports,* vol. 1, no. 4, 2019.

[18] S. Chen, Z. Kalbarczyk, J. Xu and R. K. Iyer, "A data-driven finite state machine model for analyzing security vulnerabilities," in *International Conference on Dependable Systems and Networks*, San Francisco, CA, 2003.

[19] F. Dang, H. Liang, S. Li, D. Li and H. Liu, "Design and Implementation of Computer Network Information Security Protection Based on Secure Big Data," in *IEEE 3rd IICSPI*, Chongquing City, China, 2020.

[20] F. Wang, H. Wang and L. Xue, "Research on Data Security in Big Data Cloud Computing Environment," in *IEEE 5th IAEAC*, Chongqing, China, 2021.

[21] X. Sun, P. Liu and A. Singhal, "Toward Cyberresiliency in the Context of Cloud Computing," *IEEE Security & Privacy,* vol. 16, no. 6, pp. 71-75, 2018.

[22] B. Spivey and J. Echeverria, Hadoop Security Protecting Your Big Data Platform, Sebastopol, CA: O'Reilly Media, 2015.

[23] P. J. Velthuis, "New authentication mechanism using certificates for big data analytic tools," KTH Royal Inst. of Tech., Stockholm, SE, 2017.

[24] S. Sinha, S. Gupta and A. Kumar, "Emerging Data Security Solutions in Hadoop based Systems: Vulnerabilities and Their Counermeasures," in *IEEE ICCCIS*, Greater Noida, India, 2019.

[25] "Apache Knox," Apache Software Foundation, 06 12 2020. [Online]. Available: https://knox.apache.org/.

[26] J. Longstaff and J. Noble, "Attribute based access control for big data applications by query modification," in *IEEE Second International Conference on Big Data Computing Service and Applications*, Oxford, UK, 2016.

[27] S. Oulmakhzoune, N. Cuppens-Boulahia, F. Cuppens, S. Morucci, M. Barhamgi and D. Benslimane, "Privacy query rewriting algorithm instrumented by a privacy-aware access control model," *Annals of Telecommunications,* vol. 69, no. 1-2, pp. 3-19, 2014.

[28] K. Zhang, X. Zhou, Y. Chen, X. Wang and Y. Ruan, "Sedic: privacy-aware data intensive computing on hybrid clouds," in *ACM Conf. on Computer and Comm. Security*, Chicago, IL, 2011.

[29] N. Khamphakdee, N. Benjamas and S. Saiyod, "Performance Evaluation of Big Data Technology on Designing Big Netowrk Traffic Data Analysis System," in *ISCIS*, Sapporo, Japan, 2016.

[30] G. S. Bhathal and A. Singh, "Big Data: Hadoop framework vulnerabilities, security issues and attacks," *Elsevier Array,* Vols. 1-2, p. 100002, 2019.

[31] "Apache Ranger," Apache Software Foundation, 03 09 2020. [Online]. Available: https://ranger.apache.org/.

[32] J. Wang, D. Crawl, S. Purawat, M. Nguyen and I. Altintas, "Big data provenance: Challenges, state of the art and opportunities," in *IEEE International Conf. on Big Data*, Santa Clara, CA, 2015.

[33] "Apache Atlas," Apache Sofware Foundation, 28 06 2019. [Online]. Available: https://atlas.apache.org/.

[34] NIST, Information Technology Laboratory, Computer Security Division, "Standards for Security Categorization of Federal Information and Informatoin Systems, FIPS PUB 199," NIST, U.S. Dept. of Commerce, Gaithersburg, MD, 2004.

[35] N. O. Leslie, R. E. Harang, L. P. Knachel and A. Kott, "Statistical Models for the Number of Successful Cyber Intrusions," *Journal of Defense Modeling and Simulation,* vol. 15, no. 1, pp. 49-63, 2018.

[36] Varonis, "Cybersecurity: the motivation behind cyber-hacks," Big Data Made Simple, 30 Jul. 2019. [Online].

[37] AT&T Business - Cybersecurity, "Understanding cyber attacker motivations to best apply controls," AT&T, 19 Feb. 2020. [Online].

[38] PurpleSec, "2021 Cyber Security Statistics, The Ultimate List of Stats, Data & Trends," PurpleSec LLC, 29 Apr 2021. [Online].

[39] Verizon, "Data Breach Investigations Reports (DBIR)," Verizon, New York, 2021.

[40] Deputy Assistant Secretary of Defense, Developmental Test and Evaluation (DT&E), "The Department of Defense (DoD) Cyber Table Top Guidebook," DoD, Washington, D.C., 2018.

[41] Center for Threat Informed Defense, "attack-control-framework-mappings," GitHub, Dec. 2020. [Online]. Available: https://github.com/center-for-threat-informed-defense/attack-control-framework-mappings.

[42] Common Criteria Forum, "Common Criteria for Information Technology and Security Evaluation, ISO 15408," ISO/IEC, virtual, 2017.

[43] Office of the Under Secretary of Defense, "Cybersecurity Maturity Model Certification," DoD, Washington, D.C., 2020.

[44] W. S. Humphrey, "Characterizing the software process: a maturity framework," *IEEE Software,* vol. 5, no. 2, pp. 73-79, March 1988.

[45] Forbes and IBM, "Forbes Insights Fallout The Reputational Impact of IT Risk," Forbes, Jersey CIty, NJ, 2014.