Second Step: Proposal

The proposal must clearly explain what you are doing, why you are doing it, what is new about your project, and what is the significance of your project. The proposal should include a preliminary critical review of previous related work, specific aims, how you plan to accomplish these, and a bibliography. It should also include a realistic schedule, a budget, a list of deliverables, and a discussion of any foreseeable difficulties and how you plan to overcome them. The proposal should follow the generally-accepted guidelines for computer science research proposals—for example, as described by the National Science Foundation on their web pages. We recommend that you evolve your proposal into your final report, reusing as much text as possible.

Organize the proposal as outlined below.

1. Title Page (*1 point*)
   This can be *at most* 1 page long and must the following information:
   (a) Project title;
   (b) Date;
   (c) Investigator names, affiliations, and email addresses;
   (d) Short phrase describing the general area within security of your project;
   (e) Three to five keywords;
   (f) Brief 1–3 sentence project description — distill your proposal to one focused, well-defined question;
   (g) Assignment of responsibilities to team members (if there is more than one person);
   (h) Total budget; and
   (i) Brief list of deliverables
2. Project Summary (*1 point*)
   Also *at most* one page, the executive summary is like an abstract. It summarizes the proposal. Repeat the title, date, investigator names and affiliations, and keywords on the executive summary page.
3. Motivation (*2 points*)
   Think of this as an introduction, to get the reader interested enough in the problem to want to support your research. What are you doing and why? Why is your work significant, both within your field, and to society at large? What is challenging about your proposed work? It is often useful to list concrete problems aligned with the specific aims that solve them.
4. Previous Work (*14 points*)
   Identify and critically comment on selected relevant previous work. How is your project different and better than this previous work? Do not simply list previous work. Note that you will do a full literature review later, so focus on the important previous work that is relevant to your project.
5. Specific Aims (*5 points*)
   Concretely list possible solutions to specific problems you propose to solve. The specific aims should be quite specific, and they should match elements of your motivation.
6. Plan (*2 points*)
   How will you accomplish your specific aims?
7. Deliverables (*2 points*)
   What is the output of your project — for example, will it be a project report, PowerPoint slides, a presentation, source code, a report, or something else? As you will be making a poster for the final presentation, be sure to include this.
8. Issues (*1 point*)
   What difficulties do you foresee, and how do you plan to overcome them?
9. Bibliography (*3 points*)
   List (in proper bibliographic form) all works you need to complete your project.
10. Biographical sketches of the investigator(s) (*1 point*)
11. Schedule (*2 points*)
    List a timeline of major steps toward completing your project.
12. Budget (*2 points*)
    What resources (including your own time) do you need to complete the project? Using a spreadsheet such as Excel, summarize in one simple page your direct, indirect, and total costs. These might include your time, equipment, software, travel to conferences. As your indirect costs, include a 56.5% overhead for all direct costs. (Indirect cost

is your university's overhead. Direct costs are everything else.)
For UC Davis, use the following figures (note: these figures are close to the real ones, but are undoubtedly off a bit), and assume you need funding for 2 quarters

(a) Salaries — assume yours are $4,813 per quarter;
(b) Benefits — assume yours are $63 per quarter;
(c) Travel to a conference — assume the cost is $1,250 per team member;
(d) Supplies and equipment — put whatever you think you might need;
(e) Fees and tuition — assume $12,130 per student, and there is no indirect cost for this; and
(f) Indirect costs — assume 56.5% for all costs *except* fees and tuition.

Following the amounts, explain what each is for and how it contributes to carrying out the work. The explanations can be brief.
The sponsors will not provide funding, of course. We are having you prepare a budget so you get some experience in a critical part of writing a proposal.

13. Appendix A: Research Conference (*1 point*)
Name a refereed research conference that best matches your project, and identify a recent paper from this conference that best matches your project. Please provide its DOI (or its URL if it does not have a DOI), or attach a copy of it to the proposal.
14. Broader Impact (*3 points*)
In evaluating the broader impact of the research projects, identify and comment on both the explicit and implicit benefits of the project outcomes to the society. The comments should refer to the theoretical and practical perspectives of the outcomes relative to the specific technical research area, as well as to the broader scope of cyber security domain. How does the project advance the security dimension of the particular technology? Are there any novel and interesting aspects that are considered within the project? Does the project enhance the state-of-the-art in cybersecurity? What are the practical academic benefits and what are the benefits for the public and private sector?
The comments should also consider the broader impacts on the society in general. How well does the project improve the cybersecurity awareness? Are there any aspects addressed within the project that considered from other non-technical perspectives can contribute towards a better social welfare? How the project outcomes influence the perception of the overall national security?

Example Proposal

Exploring Data Spillage in Hadoop Clusters

redacted A, Purdue University, *redacted*@purdue.edu;
redacted B, Purdue University, *redacted*@purdue.edu;
redacted C, Purdue University, *redacted*@purdue.edu

Keywords: Data Leakage, Data Spillage, Hadoop, Hadoop Cluster, Cloud, Cyberforensics

Military and other government entities perceive data spillage from Hadoop clusters as a security threat when sensitive information is introduced onto a platform considered non-sensitive. This project focuses on tracking sensitive information in HDFS after user introduction for the purpose of removing it. Using forensic analysis of the non-sensitive cluster following the introduction of a document designated as sensitive by the project team, the goal of this project is to create a procedure to remove all material designated as sensitive from the cluster after the initial load of the data.

redacted A: Hadoop cluster configuration and system evaluation *redacted C*: Cyber forensics analysis and evaluation ALL: Experimental design and data analysis *redacted B*: Data Management and recommendation of policy related controls

Total Budget: $22,892.00. Please see detail in "Budget" section below.

Project Deliverables:

1. Final Project Report: Hadoop: Understanding, Preventing, and Mitigating the Impacts of Data Spillage from Hadoop Clusters using Information Security Controls. Expected Delivery Date: 12/8/2014.
2. Project Poster: The project team will generate a poster for presentation of the project at the annual CERIAS Symposium. Expected Delivery Date: 3/24 - 3/25/2015.
3. Research Conference Presentation/Journal Paper:
   3.1. Network and Distributed System Security Symposium/International Journal of Security and Its Applications. Expected Delivery Date: 1/1/2015
   3.2. Storage Network Industry Association Data Storage Innovation Conference. Expected Delivery: April 7-9, 2015, Santa Clara, CA, USA.

Executive Summary

Data Spillage in Hadoop Clouds Project Proposal, October 1, 2014
redacted A, Purdue University, *redacted*@purdue.edu;
redacted B, Purdue University, *redacted*@purdue.edu;
redacted C, Purdue University, *redacted*@purdue.edu

Data spillage is defined in this project as a security threat when sensitive information is introduced onto a non-sensitive platform. This project will focus on user introduction of sensitive information into an non-sensitive Hadoop cluster. Using forensic analysis of the non-sensitive cluster following the introduction of a document designated as sensitive for testing purposes by the project team, the goal of this project is to create a procedure to remove all sensitive material from the cluster after the initial load of data.

Keywords: Data Leakage, Data Spillage, Hadoop, Hadoop Cluster/Cloud, Cyberforensics

Hadoop is a popular tool that supports data storage through its HDFS file system and processing in a parallel computing environments using MapReduce (in versions prior to 2.0) and YARN (in versions at or above 2.0). Hadoop clusters are a relatively inexpensive method of storing and processing large and diverse data sets; however, organizations using Hadoop should take precautions to securely handle data processed or stored by these systems. Data spillage from Hadoop clusters is an information security vulnerability that is defined in this project as the accidental or intentional introduction of tagged information into a non-sensitive Hadoop cluster. In this project we configure a new Hadoop cluster instantiation including several virtual data nodes, load a tagged document into the cluster, then find and document, through forensic analysis, the location of every piece of the tagged document throughout the cluster. Once each instance of the pieces of the sensitive document are located and documented, the pieces of the sensitive document will be erased according to Department of Justice specifications of secure erasure of data. Maximizing the availability of cluster nodes and the cluster itself while ensuring complete erasure of the sensitive data is the overall goal of this project. Nodes subjected to the optimized process will again be analyzed forensically to determine if any sensitive data can be recovered. The creation of a process to mitigate user-initiated data spillage in Hadoop clusters will have wide ranging impact not only within the United States federal government, but for any large data processor needing to quarantine sensitive information.

Motivation

Big data sets are classified by the following attributes:

- high volume (data sizes generally measured in tens of terabytes or greater)
- variety (multiple sources and data types)
- velocity (rate in which new information is added into the dataset)
- value (the utility and quality of the dataset)

Among the numerous threats to information security, data leakage in Hadoop clusters could be overlooked, yet examples of the use of data of mixed sensitivity exist in the paring of credit card information with items purchased at a retail outlet and in the introduction of mental health information on medical charts just to name two. The ability to find and correct the spillage of sensitive data into non-sensitive environments such as Hadoop clusters will only grow with the growth of electronic data. We seek to provide a means to mitigate the spread of sensitive information caused by an inability to properly locate and remove such data from Hadoop clusters in order to add to contribute to privacy protections in “big data” environments.

Previous Work

According to Intel, growth in the Hadoop market is between 50 and 60 percent annually (Mello, 2014). As an increasingly large tool in the storage and distributed processing of large data sets, the security of Hadoop and other tools based on it, is increasingly important. Data leakage, which is defined as defined as the accidental or intentional distribution of classified or private information to an unauthorized entity, is an example of a security vulnerability in Hadoop that can impact the security of information within an organization (Anjali, M. N. B., Geetanjali, M. P. R., Shivlila, M. P., Swati, M. R. S., & Kadu, 2013). The United States National Security Agency (NSA) has defined a specific instance of data leakage, called data spillage, as the transfer of classified or sensitive information into an unauthorized/undesignated compute node or memory media (e.g. disk)(Mitigations NSA Group, 2012). Requirements of federal agencies to control the distribution of sensitive information, if combined with the use Hadoop clusters for data

storage and/or processing make data spillage in Hadoop clusters a serious potential information security vulnerability within the United States government.

Xiao and Xiao (2014) discuss a process to prevent malicious actor from adding a node to a Hadoop cluster which could cause the exfiltration of information from the cluster, the failure of the cluster, or other negative impacts on cluster functionality. While the addition of a malicious node to a Hadoop cluster may be an important information security vulnerability to a subset of organizations using Hadoop, the vulnerability created through user introduction of sensitive information into a non-sensitive Hadoop cluster seems to be a more likely scenario for data loss. Assuming that introductions of sensitive data into non-sensitive Hadoop clusters are inevitable due to mis-classification of data or human error, a process for mitigation of such incidents is important. Cho, Chin, and Chung (2012) have established guidelines for locating and analyzing forensically data in an HDFS file system, but their work does not address the deletion of data from the cluster.

Because Cho, Chin, and Chung focus on Hadoop forensics, deletion of data from Hadoop clusters may be considered out of scope since it is an incident management function; however, information that exists regarding incident management in Hadoop clusters is sparse. Current articles in the trade press that discuss incident management and data security within Hadoop clusters focuses on making difficult the process of accessing the cluster and exfiltrating data from it (Garcia, 2014; Hortonworks, 2014). Therefore, our work will use established digital forensics processes to locate data within the HDFS file system and understand how data is processed within the system. Following that process, we will examine ways to completely remove data from Hadoop clusters in a way that maximizes the availability of the cluster.

Specific Aims

NSA believes that it must find each occurrence of sensitive data that has been introduced into a non-sensitive cluster, remove from the cluster all nodes on which the sensitive data resides, erase all data from the node, rebuild that node, and re-insert it into the cluster. This project will create a process for finding all occurrences of sensitive data on the cluster through forensic analysis of the nodes, and determine the optimal process for removing the sensitive data and recovering the impacted nodes. These processes will be evaluated for effectiveness and efficiency on the same Hadoop cluster on which they were created. Because the cluster used in this research is small, evaluating the effectiveness and efficiency of these processes on larger clusters is an area of future research.

The goals of the project are to:

1. Find, if possible, all the nodes on which the sensitive data resides following the upload of a document designated by the project team as sensitive.
2. Forensically examine the impacted nodes in order to understand how the data is stored and how it might be removed. Removal processes will be governed by the need for complete removal of sensitive data to NSA specifications and maximization of the availability of impacted cluster nodes. Preservation of non-sensitive data is considered of lower importance than complete removal of the sensitive information, and availability of the cluster.
3. Forensically examine the impacted nodes following the sensitive data removal data process in order to confirm the inability of potential attackers to find residual sensitive information.
4. To the extent possible, automate the above listed process

Plan

Assumptions:

1. The Hadoop Cluster is an unclassified silo running Hadoop version 2.0.
2. A user loads non-sensitive data into the system.
3. The processes involved in this project will work similarly in physical and virtual environments.
4. Selected nodes will be a virtual disk and the virtual disk image will be used for forensic processes based on uniqueness of data storage methods on the nodes.

Goal: Detection of the introduction of sensitive data into a non-sensitive Hadoop cluster prior to processing by YARN.

1. Build a small Hadoop version 2.0 cluster containing one name node and no more than 30 data nodes.
2. Introduce a document designated as sensitive into the Hadoop cluster to simulate the introduction of classified or sensitive document onto an unclassified or non-sensitive Hadoop cluster.
3. Use data from the name node, and if necessary, forensic processes on the data nodes to locate on the cluster all artifacts of the designated sensitive document.

4. Forensically analyze one of the data nodes in order to determine how the sensitive data is stored, and how to remove that data completely while minimizing the impact on the availability of impacted nodes and on the Hadoop cluster.

Data Management Plan

Managing data throughout the research lifecycle is essential to this project in order for it to be replicated for further testing, as guidance for various federal agencies, and for re-testing as new versions of Hadoop are released. The ability to publish and re-create this work will involve the thorough documentation of processes and configuration settings. Initial data management will document the initial hardware and software specifications of the hardware and software used, the configuration processes used in configuring both the Hadoop cluster and the physical and virtual nodes, and a detailed description of the tagged data and the process of loading it into the cluster. The setup of the environment will be documented as a set of instructions that will be uploaded to PURR. These setup instructions will be verified by having another team member duplicate the environmental configuration upon completion of the final report. Software or data integral to the reproduction of the experimental aspects of this study will be backed up to PURR when such backups are technically feasible and will not violate applicable laws. Processes and results of forensic analysis will also be thoroughly documented, verified by a second team member on additional nodes, and uploaded to PURR.

PURR will serve the data management requirements for this project well because it provides work flows and tools for uploading of project data, communication of that data, and other services such as: data security, fidelity, backup, and mirroring. Purdue Libraries also offers at no cost to project researchers consulting services in order to facilitate selection and uploading of data, inclusive of the generating application and necessary metadata that will ensure proper long-term data stewardship. Documentation of results and processes will also be uploaded to the team's shared space on Google Docs as a backup to PURR. The contact information of the Purdue Libraries data manager assigned to this project, will also be uploaded to PURR to facilitate long term access to, and preservation of project data.

Deliverables

1. *Final Project Report*: Hadoop: Understanding, Preventing, and Mitigating the impacts of Data Spillage from Hadoop Clusters using Information Security Controls. This report will contain: a review of the relevant literature, results of the study, a recommended process for mitigation of the introduction of sensitive data into a Hadoop 2.0 cluster, and suggestions for future work. Expected Delivery Date: 12/8/2014.
2. *Archived Project Artifacts*: All experimental results and project artifacts that facilitate the re-creation of those results will be archived on PURR as described in the section of this report entitled, "Data Management Plan." Expected Delivery Date: 12/8/2014
3. *Weekly Project Updates*: Project stakeholders will be updated each Thursday on the progress of the project through the completion and uploading to PURR of a progress report through November, 2014.
4. *Project Poster*: The project team will generate a poster for presentation of the project at the annual CERIAS Symposium. Expected Delivery Date: 3/14/2014.
5. *Research Conference/Journal Paper*:
   5.1. Network and Distributed System Security Symposium/International Journal of Security and Its Applications.
   5.2. Storage Network Industry Association Data Storage Innovation Conference. Expected Presentation: April 7–9, 2015, Santa Clara, CA, USA.

Issues

Understanding data processing in the Hadoop environment well enough to mitigate the introduction of sensitive information onto a non-sensitive cluster presents the following challenges:

1. *Virtualization*: In order to increase the speed of setup and cost effectiveness of the Hadoop cluster, the nodes in this project will be run on virtual machines. We do not expect, but must consider the potential that virtualization of the Hadoop environment may create a material impact on the results of this study. These concerns are especially prominent in aspects of forensics because the forensic processes for virtual machines is substantively different than those for physical computing resources.
2. *Administrative Support*: The Hadoop cluster and its virtual environment will be administered to by the researchers on this project. Project researchers are capable of handling system issues surrounding the virtual environment and Hadoop clusters, but major system issues such as loss of a portion of the cluster and power backup are outside of the scope and expertise of the project team. Because these issues are outside of the control of team members, they would cause major delays in project completion if experienced.

3. *Time*: This project is being undertaken as a project for a course lasting 16 weeks. Limiting the duration of the project to 16 weeks impacts the scope of the project, and increases the likelihood that deviations to the project plan could cause a major impact on project deliverables.

Bibliography

1. Anjali, M. N. B., Geetanjali, M. P. R., Shivlila, M. P., Swati, M. R. S., & Kadu, N. B. (2013). Data leakage detection. In International Journal of Computer Science and Mobile Computing
2. Cho, C., Chin, S., & Chung, K. S. (2012). Cyber Forensic for Hadoop based Cloud System, 6(3), 83–90.
3. Garcia, E. (2014). Bigger Data, Smaller Problems: Taking Hadoop Security to the Next Level — SecurityWeek.Com. Retrieved October 10, 2014, from http://www.securityweek.com/bigger-data-smaller-problems-taking-hadoop-security-next-level
4. Hortonworks. (2014). Hortonworks Acquires XA Secure to Provide Comprehensive Security for Enterprise Hadoop - Hortonworks. Retrieved October 10, 2014, from http://hortonworks.com/blog/hortonworks-acquires-xasecure-to-provide-comprehensive-security-for-enterprise-hadoop/
5. Mello, J. (2014). Rapid Growth Spurs Competition in Hadoop Market. Retrieved October 09, 2014, from http://data-informed.com/rapid-growth-spurs-competition-hadoop-market/
6. NSA Mitigations Group. (2012). Securing Data and Handling Spillage Events [White Paper]. Retrieved from https://www.nsa.gov/ia/_files/factsheets/final_data_spill.pdf
7. Xiao, Z., & Xiao, Y. (2014). Achieving Accountable MapReduce in cloud computing. Future Generation Computer Systems, 30, 1–13. doi:10.1016/j.future.2013.07.001

Biographical sketches of the team members

Redacted A is a doctoral research assistant for Dr. Dark. Her research background is in system modeling and analysis that allows her to approach this project from a statistical-based modeling approach. Over the summer, she worked on developing a data motion power consumption model for multicore systems for the Department of Energy (DOE). In addition, she is working towards growing expertise and experience in big data analytics under her adviser Dr. Springer, the head of the Discovery Advancements Through Analytics (D.A.T.A.) Lab.

Redacted B is a Ph.D. student specializing in secure data structures for national health data infrastructure. His most recent experience includes security and privacy analysis of laws, policies, and information technologies for the Office of the National Coordinator for Health Information Technology within the Department of Health and Human Services.

Redacted C is currently pursuing his second masters at Purdue University in Computer Science and Information Technology. *Redacted C* is a qualified professional with diverse interests and experiences. Prior to his graduate studies at Warwick University, *Redacted C* spent four years studying computer science engineering at MVGR College of Engineering. Continuing his education, he graduated with a master in Cyber-security and management degree from University of Warwick, UK. While at Warwick, *Redacted C* spent his days building a prototype tool for HP cloud compliance automation tool semantic analysis phase at Hewlett Packard Cloud and Security labs. Further, *Redacted C* worked at Purdue University as a Research Scholar researching in the domain of cyber security and digital forensics. He is an active member of IEEE.

Schedule

Analysis of Data Spillage in Hadoop Clusters Project Activities Schedule

	Project Goal Ref #	Timeline				
Activity		October 1-15	October 16-31	November 1-15	November 16-30	December 1-15
Final Approval of Project Scope: 10/1	1					
Hadoop Cluster Configuration: 9/22-10/10	1					
Data Load: 10/15	1					
Analysis and Documentation of Data Locations: 10/17-10/31	1					
Imaging and Foresnsic Analysis of Selected Nodes: 11/1-11/7	2					
DoJ Specification Data Removal: 11/7-/11/8	3					
Forensic Analysis of DoJ Processed Disks: 11/10-11/14	3					
Optimize Process Find – Reintroduction of Node: 10/15-12/1	4					
Final Report of Results and Poster Creation: 11/14-12/5	4					

Green cells indicate ongoing work. Red cells indicate completion of work.

Budget

Item	Unit	Cost/Unit	Total
A.1. Team Hours	3 members, 12 weeks	$1,984/member	$5,592.00
C. Fringe Benefits	3 members	$750/member	$2,250.00
E. Conference Travel	3	$2,194.67	$6,584.00
I. Purdue Indirect	1	$15,677.86 * .54	$8,466.00
TOTAL			**$22,892.00**

A.1. Senior Personnel:
All three senior personnel will work for 3 months (12 weeks) on the project at an hourly rate of $16.534/hr (based on SFS monthly stipend).
3 months * 40hrs/month * $16.534/hr = $1984

C. Fringe Benefits:
A total of $750 is requested to cover fringe benefits, corresponding to 20% of pay rate to cover medical, vision and dental insurance.

E. Travel:
All three senior personnel will attend the Storage Network Industry Association Data Storage Innovation Conference. April 7–9, 2015, Santa Clara, CA, USA.
Estimated full conference academic pass = $695/attendee * 3 attendees = $2084.
Estimated travel expenses (hotel, transportation, meals, others) = $1500/attendee * 3 attendees = $4500.

I. Indirect Costs:
The indirect rate for this project is 54% according to requirements of Purdue University
$15,677.86 * 0.54 = $8466

Broader Impact

As usage of large data sets continues to grow, opportunities and needs to process and use sensitive and public data together will also grow. Examples of the use of data of mixed sensitivity exist in the paring of credit card information with items purchased at a retail outlet and in the introduction of mental health information on medical charts just to name two. The ability to find and correct the spillage of sensitive data into non-sensitive environments such as Hadoop clusters will only grow with the growth of electronic data. This project will produce a process that can not only be used to mitigate the introduction of sensitive material into non-sensitive Hadoop clusters, but could be used as the basis for such processes in other distributed data storage and processing systems. In either the Hadoop context, or more broadly, the procedure detailed by this project could potentially add to the electronic protections of privacy for billions of people worldwide.