https://doi.org/10.48009/4_iis_2015_121-123

Issues in Information Systems
Volume 16, Issue IV, pp. 121-123, 2015

A SECURITY VULNERABILITIES ANALYSIS OF THE APACHE BIG DATA STACK AND HIGH PERFORMANCE COMPUTING TECHNOLOGIES

James D. Sissom, Southern Illinois University Carbondale, jsissom@siu.edu
Alex M. Glasnovich, Southern Illinois University Carbondale, lilglass.alex@siu.edu

ABSTRACT

The term big data relates to the massive amount of many different types of data that is produced very quickly from a large number of sources. Real-time datasets require a variety of tools that link data to powerful processors and software. This is fertile ground for threat actors that act brazen because of their increasing success and experience in an ever evolving data security arena where it may take professionals an average of 11 years of security experience to acquire the skills to defend against modern day attacks. [1]

Keywords: Apache Big Data Stack (ABDS), Common Vulnerabilities and Exposures (CVE), Common Vulnerability Scoring System (CVSS), High-Performance Computing (HPC), and National Vulnerability Database (NVD).

INTRODUCTION

Both commercial and research organizations are rapidly expanding data science capabilities depending on the ability to analyze and compute large amounts of data. Data is integral to every major trend and is impossible to completely lock down without threatening to ruin a brand's reputation or very existence of a key project. The significance of data in commercial sectors and research fields allows productivity and the creation of knowledge from large and complex volumes of data. In addition, increasingly visible cyber warfare is raging bringing the combatant's tactical and strategic language and landscape out into the open. Early 2014 exposed our brittle infrastructure with OpenSSL Heartbleed vulnerability (CVE-2014-0160*)* [1] exposing the memory of computer systems. Technological advancement creates a duality where highly-skilled data science actors and threat actors engage in digital Darwinism; how can big data survive the evolving threat?

High-Performance Computing (HPC) and the Apache Big Data Stack (ABDS)

At the ORAU Big Data conference in March 2015, Dr. Geoffrey Fox, [2] confirmed digital data reached 9 zettabytes. An IDC Forecast predicted digital data will grow 2.75 zettabytes in 2012 and reach nearly 8 zettabytes by 2015 [3]. This number only includes consumer data thus it does not include scientific data. Significant, because this target was reached much faster than predicted. Dr. Fox discussed his big data research in his presentation. He also discussed a table called, "Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies" [4] that currently lists 21 layers and over 350 software packages. Qiu, et al. [5] provides a discussion of the importance of advanced analytics in the enterprise and scientific communities:

The importance of advanced analytics to derive insight and knowledge from increasing volumes of complex data will continue to grow. The enterprise community has made impressive gains and seem to have converged around the Apache stack, a distinctive feature is the existence of many implementations of the specific components of the Apache stack, providing sufficient richness in the trade-off between performance and capability. In contrast, within the scientific computing community, progress has been reliant either on long-term foundational advances or short-term hardware fixes as opposed to integrated approaches that marry the relative technical strengths of the two communities yet deliver these as implementations usable on high performance and distributed computing HPDC infrastructure such as XSEDE, OSG and other domain-specific infrastructure.

Based on a review of the Kaleidoscope of ABDS and HPC technologies, a gap exists between the enterprise (commercial) community and the scientific computing community. Fox and fellow research colleagues are developing "*HPC-ABDS will utilize and expose the integrated relative technical strengths of the two hitherto disjoint approaches and communities, yet it will focus on delivering these as production grade implementations that*

will bring the best-of-both to shared-infrastructure" [5] this paper is using an analysis of the CVE database as one approach to review how well big data software communities are discovering and disseminating potential security vulnerabilities.

RESEARCH METHODOLOGY

Our research question is simple: among the most current list of the 21 Kaleidoscope layers and ~350 software packages (as of May 15, 2015), what is the total number of findings that will match a keyword search using the CVE search engine? How many searches will render a recent high severity in the Common Vulnerability Scoring System (CVSS)? Researchers will use the CVE search engine found at the following URL: https://web.nvd.nist.gov/view/vuln/search

Results

The keyword search for each software placed into the CVE search engine resulted in a total of 17,707 findings, shown in Table 1. Table results are broken down into five Kaleidoscope Software Package Layers.

Table 2 results are broken down using the five Kaleidoscope Software Package Layers and displays the number of searches that rendered a recent high severity using the CVSS classification. Researchers utilized the search feature, "Search Last 3 Years" to examine and record the most recent CVSS high severity findings.

Table 1. National Vulnerability Database Search Results (Total Number of CVE Findings)

Kaleidoscope Software Package Layers	**Total Number of CVE Findings**
Cross-Cutting Functions 1 - 4 CVE Findings	2,024
Layers 5 - 9 Total CVE Findings	501
Layers 10 - 12 Total CVE Findings	15,019
Layers 13 - 15 Total CVE Findings	153
Layers 16 - 17 Total CVE Findings	10
	17,707

Table 2. National Vulnerability Database Search Results (High Severity Findings)

Kaleidoscope Software Package Layers	**CVSS High Severity Findings**
Cross-Cutting Functions 1 - 4 CVE Findings	279
Layers 5 - 9 Total CVE Findings	80
Layers 10 - 12 Total CVE Findings	3,977
Layers 13 - 15 Total CVE Findings	40
Layers 16 - 17 Total CVE Findings	6
	4,382

NVD Search Resource Status

The NVD database resource status is displayed in Table 3, the table contents displays the complete content type and number of CVE Vulnerabilities, the National Checklist Program (US-CERT), the Open Vulnerability Assessment Language (OVAL) queries, and Official Common Platform Enumeration (CPE) names.

Table 3. National Vulnerability Database Resource Status (as of 7/7/2015)

CVE Vulnerabilities	Checklists	US-CERT Alerts	US-CERT Vuln Notes	OVAL Queries	CPE Names
71,051	298	249	4,367	10,286	104,927

CONCLUSIONS

It is clear from the CVE search engine results that a significant number of findings amongst the ABDS-HPC requires diligence and continued participation from the CVE Numbering Authorities (CNAs) [6]. Based upon the results of this study, we conclude that significant, recent, high-level, security exploits exist in the HPC-ABDS stack. Public and private big data vendors should continue to work closely with the CVE project in order to assist with the analysis, research, and processing of incoming vulnerability submissions. The task of discovery, matching to known issues, numbering, and refining the submission to the final CVE content requires deep research. The CVSS severity, CVE numbering method were matched during this research and left out of this paper. Additional information regarding the types of CVE identifiers and specific threats to Big Data platforms will be complete and ready to distribute, display, and discuss at the International Conference.

REFERENCES

1. Websense 2015 Threat Report. (2015) https://www.websense.com/assets/reports/report-2015-threat-report-en.pdf. Published April 8, 2015. Downloaded on May 1, 2015. Produced by Websense Security Labs.
2. Geoffrey Fox. (2015, March). Classification of Big Data Applications and Implications for the Algorithms and Software Needed for Scalable Data Analytics. *70th Annual Meeting of the ORAU Council of Sponsoring Institutions*. Oak Ridge, Tennessee.
3. Vesset, D., Woo, B., Morris, H.D., Villars, R.L., et al. IDC Worldwide Big Data Technology and Services 2012-2015 Forecast. (2012, March). #233485. Retrieved from http://siliconangle.com/files/2012/05/Where-is-your-data-FINAL-5a.png
4. Fox, G., Qiu, J., Jha, S., Kamburugamuve, S., and Luckow, A. (2015) Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies. Retrieved from http://hpc-abds.org/kaleidoscope
5. Qiu, J., Jha, S., Luckow, A., and Fox, G.C. (2014) Towards HPC-ABDS: An Initial High-Performance Big Data Stack to be published in proceedings of *1st Big Data Interoperability Framework Workshop: Building Robust Big Data Ecosystem ISO/IEC JTC 1 Study Group on Big Data.* March 18 - 21, 2014 San Diego Supercomputer Center, San Diego, CA, USA.
6. Common Vulnerabilities and Exposures Numbering Authorities. (2015, July 8). Retrieved from https://cve.mitre.org/cve/cna.html