HIGH PERFORMANCE DATA ANALYTICS

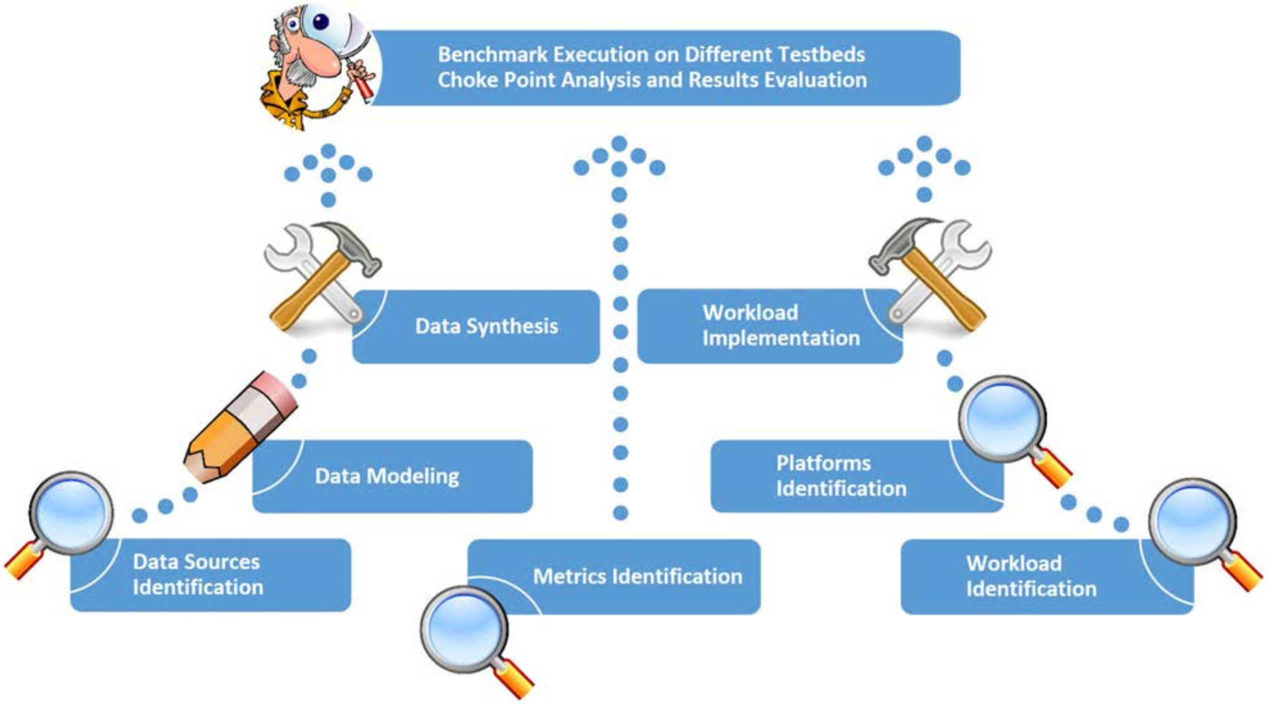
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Big Data Benchmarking Suite for Cybersecurity Analytics

CHALLENGE

Develop a benchmark for evaluating performance and bottlenecks of big data systems with a focus on cybersecurity-related workloads and data sets.

Data collection and analysis are rapidly changing the way scientific, national security, and business communities operate. Data analytics, especially graph analytics, have received much attention in the last 50 years. Moreover, with this increasing interest in graph processing, the diversity of graph data sets and graph processing algorithms has also increased. As a consequence, various graph processing platforms have been developed that combine the hardware and software for completing a graph processing task. In spite of their practical applications success, it is difficult to make deployment decisions among the large variety of platforms because of the lack of comprehensive understanding of their



Cybersecurity-specific benchmarking methodology for big data systems.

performance. Before system designers, programmers, and researchers within the cybersecurity domain can optimize the performance, resilience, and energy efficiency of these systems, application-specific benchmarking is mandatory.

CURRENT PRACTICE

Over the years, several studies employing big data benchmarks have been conducted on big data systems and architectures. However, state-of-the-art big data benchmarks exhibit two main characteristics: 1) most use data sets built for the application domains of search engines, social networks, and e-commerce, and 2) they usually implement a generic workload (e.g., the breadth-first search [BFS] traversal implemented in the de facto standard benchmark Graph 500) to embrace the widest possible range of applications. Because neither the data sets nor the workloads represent the cybersecurity domain, their results cannot be successfully generalized in that context. The latter mandates using application-specific benchmarks with application-specific data generators that are able to produce an arbitrary amount of data with high veracity property.

To represent the workload perspective, it must include typical operations executed in the cybersecurity domain, such as queries on nodes, edges, paths, and subgraphs.

TECHNICAL APPROACH

We propose to design and employ an exhaustive benchmarking methodology for the efficient and reliable generation of a benchmark to evaluate the performance of graph processing systems with respect to the cybersecurity application domain. This task will be achieved by addressing significant challenges faced by current benchmarks used to evaluate graph processing systems. First, we will identify data models on the basis of real-world data pertaining to cybersecurity applications to generate synthetic data that are used as inputs of workloads. After achieving that first goal, we will compare and evaluate the conformity of the synthetic data with the inherent and important characteristics of raw real-world data by exploring various known data veracity evaluation metrics, such as, literal diversity, relationship specialty, data set coherence, and others. The next step will be to identify and implement application-specific query workloads on different software platforms, making sure they accommodate complex scenarios that will enable assessment of the system performance and any bottlenecks.

IMPACT

Using application-specific benchmarking can lead to an efficient and accurate evaluation of big data systems, especially in the cybersecurity domain. By exploring and testing various data models for synthetic and real data sets with data veracity evaluation metrics and generating application-specific workloads, this project will be particularly important to the HPDA projects requiring cybersecurity-related data or workloads to promote and test developed theories, algorithms, and technologies.

Contacts

Stefano Iannucci
Principal Investigator
(662) 325-0912
stefano@dasi.msstate.edu

David Dampier
Director, Distributed Analytics
and Security Institute
(662) 325-0779
dampier@dasi.msstate.edu

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

MISSISSIPPI STATE
UNIVERSITY