Managing Large-Scale Heterogeneous Deployments for Cybersecurity

J.S. Hurley
ODNI, Bethesda, USA

john.s.hurley@odni.gov

Abstract: In cyber defense, we must contend with the massive amounts of data being generated in a variety of different formats and speeds. Unfortunately, traditional tools and methods are not meeting the requirements for scale and speed and rely too heavily on heuristics. Advancements in mobile technologies and the Internet of Things (IoTs) will continue to contribute to the additional growth in data volumes anticipated for the foreseeable future. As data continues to grow in complexity and scale, cyber professionals must rely upon models that are more elaborate and sophisticated to predict future behavior. More complex models can give additional inference capabilities; however, they are also difficult to scale and deploy in real-time environments. Managing large-scale, heterogeneous deployments for cybersecurity is challenging. Hardware capabilities and software tools both motivate and limit computational and inferential objectives. Hence, the interplay between data science (especially machine learning) and computation become more significant than ever to explore to gain more insight into heterogeneous deployments and how they can be more effectively managed. In this study, we identify ways in which data science tools and techniques can be used in improving the management of large-scale heterogeneous deployments for cybersecurity.

Key Words: Heterogeneous Deployments, Data Science, Heuristics, Inference, Cybersecurity.

1. Introduction

A large-scale set of heterogeneous devices reflects a fundamentally complex system with unique requirements and challenges driven by diverse devices with different operating systems and/or protocols. These deployments are complex, not only because of the need to integrate different languages, platforms, users, and technologies, but to do so seamlessly, with limited or no disruption of services. Other key requirements include the need to have continuous, resilient operations and high availability. Deployments must be flexible and dynamic enough to accommodate applications that can run in any environment. The ability to deploy at the edge, on-premises, or in the cloud requires diverse operation and security requirements that can complicate deployment. The advancements of technology, especially in the areas of mobile technologies and the Internet of Things (IoTs), demands that environments be readily scalable to accommodate future data growth. Similarly, network operators should anticipate that new systems and platforms will expand system diversity as emerging technologies and novel devices are incorporated into already complex systems.

It is critical to understand the layout of how devices and nodes are physically or logically connected and to understand how data will move through the network environment and where particular chokepoints or potential points of obstruction could exist within a network. An understanding of the network topology is a primary means of establishing effective network management and monitoring. No single network topology approach is perfect or inherently better than any other.

From a security and operations standpoint, it is important to understand and prioritize applications and the data associated with them. A consistent problem that has plagued organizations is the inability or unwillingness to prioritize network infrastructure and data management, including even keeping appropriate records of the location of critical data. Such behavior has unnecessarily complicated the ability of organizations to assign their limited resources to the most mission critical information assets which should have the most stringent security and operations requirements. Consistently, organizations treat all applications with the same value leading to unnecessary waste of funds, especially in terms of storage and security requirements. All applications and data should not be treated the same, and the most mission critical applications and data should receive the most attention.

Scalability and security are dominant issues in heterogeneous deployment. Network virtualization can play an important role in both, as strategies are increasingly being employed to improve heterogeneous deployment methods. Virtualization provides many network benefits, especially by allowing multiple isolated virtual networks to share the same underlying physical infrastructure. Virtualization strategies that virtualize computing, processing, and networking enable virtual networks to be added and scaled and allow networks to be spun up more quickly in response to shifting business requirements. Faster service delivery, improved control,

and enhanced operational efficiency are all byproducts of the flexibility that network virtualization allows (English, 2022). However, a downside of the virtualization is the network sprawl created when network administrators overindulge in the creation of virtual networks. In these cases, excessive resource consumption and network complexity can result. In addition, as enterprises migrate increasingly to virtual networks, the impact of new architectures on resource consumption, resilience and security must be given consideration.

In this study, data science tools and technologies will be used to address some of the challenges associated with managing heterogeneous deployments. The study will begin with a discussion of the different nodes, devices, and links and the importance of knowing how they are connected physically and logically. Next, the focus is on the issues of application and data prioritization and their importance in terms of where the limited resources available to an organization should be allocated, especially for security purposes. Next, the emphasis is specifically on the challenges associated with each of the computing environments. In the following section, considerations will be given to different data science tools and technologies and how they may alleviate some of the problems associated with the unique computing environments. Specifically, the focus will be on identifying and prioritizing the appropriate applications and data, then setting up a hierarchy of mission critical applications and associated data, wherein the data associated with the level of priority is placed in the appropriate storage locations. Next, the computing environments (cloud, edge, and on-premises) that are best suited for the selected applications will be coordinated. Lastly, overall recommendations that may mitigate some of the current challenges for deployment will be provided.

2. Network Topology

Network configuration or topology is key to determining network performance and provides the confidence to securely operate within the network. Information Technology (IT) and security professionals must have a keen understanding of how the different nodes, devices and links on networks are logically related and physically arranged to optimize network performance while mitigating network attacks. There are several options available to professionals with the usual constraints, i.e., size, scale, budget, mission, and goals of the organization. As expected, there are advantages and disadvantages to each option, but certain arrangements can provide a greater degree of security and connectivity. In physical network topology, there is greater emphasis on the physical connections and interconnections between the nodes and the network. The concept of logical network technology is a bit more abstract, referring to how the network is arranged as well as how data moves through the network. Both the physical and logical approaches to network technology play important roles in application and network performance.

3. Application/Data Prioritization

Organizations consistently engage in practices that adversely impact data, system and network performance and security. There is a general lack of knowledge of the level of priority of the applications and data for which an organization is responsible. In other words, applications, and data, for the most part are lumped into the same barrel with little to no distinction between those that are and are not mission critical. In addition, there is a general unawareness of what data the organization possesses; where the data is located; and how the data should be grouped or categorized. As a result, organizations tend to treat all data or at least most of the data with the same level of priority—an unnecessary use and cost of data storage and security resources. Not all data is created equally. It is more reasonable and practical for an organization to decide on the data that is sensitive or mission critical. Next, less critical data should also be assigned at the appropriate levels. All of this begins with first identifying and cataloguing the data to determine the types of data that the organization possesses. Finally, it is necessary to determine which individuals, groups, departments, and partners need certain data, and for what length of time.

4. Cloud, Edge, On-Premises Environments

We are familiar with on-premises and cloud computer environments. On-premises simply refers to an environment in which resources are deployed in-house and within an enterprise's local IT infrastructure. The services are generally privately owned and controlled locally. In a cloud environment, the resources are generally hosted on a vendor's server and not physically located in a local facility. In an edge computing environment, applications are brought closer to the data sources, which has numerous advantages, including improved response times and better bandwidth availability. A closer examination of the three different environments is made with distinct advantages and disadvantages of each being provided below.

An On-premises environment consists of IT infrastructure software and hardware applications that are hosted on-site. In spite of the local management and access to resources, on-premises environments face several challenges, including peak-capacity planning, continuous upgrades, reliability and runtime, high-availability, and disaster recovery. A major concern for an on-premises computer-aided design (CAD) infrastructure is reliability. Depending on the tracking methods, tracking, and calculating 100% uptime can be imprecise *(Varde, et al., 2021)*. On-Premises computing is about addressing computing requirements on the site of the organization. The biggest advantage likely for on-premises computing is security and data protection, since the data is stored locally on site, so there is full control over the data and the security. Locally stored data can still be accessed on the premises even when external network or Internet connections are interrupted. Upfront costs can be very high given that the functionality of the programs and back up of the data must be ensured. Typically, hardware requirements must be met by on-site devices because the software runs locally which can contribute to the costs of maintaining the right hardware. Scalability of on-premises environments remains a challenge from a cost and infrastructure reframing standpoint *(Kemper, 2021)*.

In cloud computing, the focus is on the on-demand access to services and resources via the Internet that are usually hosted at a remote data center that is managed by a cloud service provider (Srivastava & Khan, 2023). *For cloud computing, there are usually three distinct categories defined including: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Some of the advantages of cloud computing are Reliability, Strategic Edge, Security (depending on who you ask. Security has been improved, but there is still a high reluctance depending upon the cloud environment chosen. Private and hybrid cloud environments due to compliance requirements for certain organizations and applications are the only options available* (Duggal, 2022). A core component of the cloud-native computing environment, microservices, is being widely embraced as a very new and popular approach for developing and deploying cloud applications that require higher levels of agility, scalability, and reliability. A microservice-based cloud application architecture advocates decomposition of monolithic application components into "microservices" – independent software components. Unfortunately, this also creates complex runtime performance monitoring and management challenges because the independent microservices can be developed, deployed, and updated independently of each other (Noor, et al., 2019). The strength of the microservice model (agility, independence, diversity) also presents cybersecurity and monitoring challenges.

Edge devices are associated with hardware that controls data flow at the boundary between two networks. These devices can function in a variety of different ways including filtering, monitoring, processing, routing, storing, translating, and transmitting data between networks. In edge computing, processes are decentralized and occur in a more logical geographic location. The IoTs and cloud computing have raised the profile of edge devices, echoing the need for more advanced services, computing power and intelligence at the edge of the network. Unfortunately, Edge devices can increase the risk of cybersecurity threats to an organization's network. As might be expected, the deployment of hundreds of edge computing devices creates hundreds of potential entry points for Distribute Denial of Service (DDoS) attacks and other security breaches. The level of vulnerabilities is especially concerning from a security perspective because many of the endpoints are "smart," i.e., features are built-in for Internet connectivity. In addition, IT no longer has full visibility or centralized control. CISOs and CTOs are then presented with the significant security challenge of protecting the data that moves through or resides in edge devices, as the attack surface expands *(Ray, 2022)*. The capacity of edge devices is much more modest than that of a mini-cloud, and as a result requires new algorithms, methods, and policies to set the proper virtualization strategies for edge devices. Edge computing enables data to be analyzed locally, i.e., closer to where the data resides, in or near real time with minimum latency. Some advantages of edge computing are high speed, reduced latency, better reliability for faster content delivery and data processing. Edge computing offers a far less expense for versatility and scalability. Some of the disadvantages of edge computing are more storage capacity required, security challenges due to high amount of data. Edge computing also requires advanced infrastructure *(Arora, 2022)*.

5. Results and Discussions

In this study, we seek to address a number of the critical issues that plague organizations in terms of their applications and data in heterogeneous deployments. It is concerning how few organizations can lay (digital) hands on or even describe the location of their critical data when needed. Equally disturbing is the unawareness of how the data is being accessed and with whom it is shared. Most organizations still have

trouble addressing these issues and as such are significantly challenged by new global compliance requirements that mandate the understanding of the location of sensitive data as well as the level of security and protection of that data. The fact that data growth is expected to continue in the foreseeable future suggests the problem will only get worse as time goes on. With growing requirements and limited resources, we suggest a nuanced approach where all data is not treated the same. Given the limited resources within our organizations, it is critical that we establish hierarchies from the highest to the least critical data. It is evident that data that is critical to the Mission of the organization should receive the highest priorities. Hence, within this study, the attempt is made to understand where applications and data can be located so that data can be accessed more quickly. In this study, there are ten applications identified from E1-E10, with each having an arbitrary focus area that ultimately decides their position due to how critical the application and data are to the Mission of the organization, see Table 1. The requirements for the applications were generated randomly. The data for each of the applications is stored in a specific storage location, similarly, identified as storage data location (SDLn) from S1-S10. The applications are later ranked in terms of how critical they are to the Mission of the organization from High to Low, see Table 2.

Table 1: Applications, Focus Areas, Mission Critical Status, Storage Data Locatio

AppName	Focus Area	MCS	SDLn
E1	FA1	Medium	S1
E2	FA2	Low	S2
E3	FA3	High	S3
E4	FA4	Low	S4
E5	FA5	Medium	S5
E6	FA6	Medium	S6
E7	FA7	High	S7
E8	FA8	High	S8
E9	FA9	Low	S9
E10	FA10	Medium	S10

Table 2. Applications, Focus Areas, Mission Critical Status, Storage Data Location—Prioritized by MCS

AppName	Focus Area	MCS	SDLn
E3	FA3	High	S3
E7	FA7	High	S7
E8	FA8	High	S8
E1	FA1	Medium	S1
E5	FA5	Medium	S5
E6	FA6	Medium	S6
E10	FA10	Medium	S10
E2	FA2	Low	S2
E4	FA4	Low	S4
E9	FA9	Low	S9

In Tables 3. And 4., the emphasis is placed on criteria for the three different computing environments of note, including the cloud, edge, and on-premises environment. In Table 4., the environment criteria in Table 3. is converted from categorical to quantitative representations to employ in computing program.

Table 3. Computing Environments Criteria.

Cloud	Edge	On-Premises
Little to No Upfront Costs	Most Cost-Effective	Large Upfront Costs
Data Ownership Not Transparent	Better Data Sovereignty than Cloud	Complete Control of Data
High Latency	Lowest Latency and Congestion Solution	Low Latency
Some Cybersecurity Concerns	Potential Data Privacy and Cybersecurity Concerns	Stronger Control of Security
Lower Than On-Premises Real Time Performance	Highest Real Time Performance Advantage	High Real Time Performance Advantage
Highest Scalability	Higher Scalability than On-Premises	Least Flexibility
Highest Bandwidth Flexible	Higher Bandwidth Flexibility than On-Premises	Bandwidth Upgrade Capability Lowest
Lowest Control Host, Management, and Some Maintenance	Higher Host, Management, Maintenance Control than Cloud	Highest Level of Control
Reliable, but Internet Connection Critical	Higher Liability than Cloud	Highest Reliability
Database or Performance-Intensive Apps	Greater Preferred than Cloud	Most Preferred

Table 4. Computer Environments Criteria (Numerical Conversion)

Requirements	Cloud	Edge	On-Premises
Costs	2	3	1
DO	1	2	3
Lat	1	3	2
Cybersec	1	2	3
RTP	1	3	2
Scale	3	2	1
BW	3	2	1
HMM	1	2	3
Rel	1	2	3
Apps_PI	1	2	3

Legend: Costs = Costs; DO=Data Ownership; Lat=Latency; Cybersec=Cybersecurity; RTP=Real-Time Performance; Scale=Scalability; BW=Bandwidth; HMM=Host, Management, and Maintenance; Rel=Reliability; Apps_PI=Performance-intensive Apps. 1º lowest value; 2º medium value; 3º highest value.

In Table 5., a collective arrangement of the apps, criteria, and computing environments are listed. In Figures 1., 2., and 3. correlation and heatmap plots for the applications, criteria and computing environments are presented, respectively.

Table 5. Full Apps and Computer Environments

Apps	Costs	DO	Lat	Cybersec	RTP	Scal	BW	HMM	Rel	Apps_PI	Cloud	Edge	On_Premises
1	3	3	3	1	3	1	3	2	1	2	2	3	1
2	2	2	2	3	1	3	3	3	3	1	1	2	3
3	2	3	2	2	3	3	3	1	1	1	1	3	2
4	2	3	1	2	3	1	3	1	1	3	1	2	3
5	3	3	2	1	1	2	2	2	2	3	1	3	2
6	1	2	2	1	2	1	1	2	3	3	3	2	1
7	2	1	2	1	3	1	2	1	2	3	3	2	1
8	2	1	2	3	2	1	3	2	2	1	1	2	3
9	2	3	2	2	2	1	1	1	3	2	1	2	3
10	3	1	2	3	1	3	3	3	1	2	1	2	3

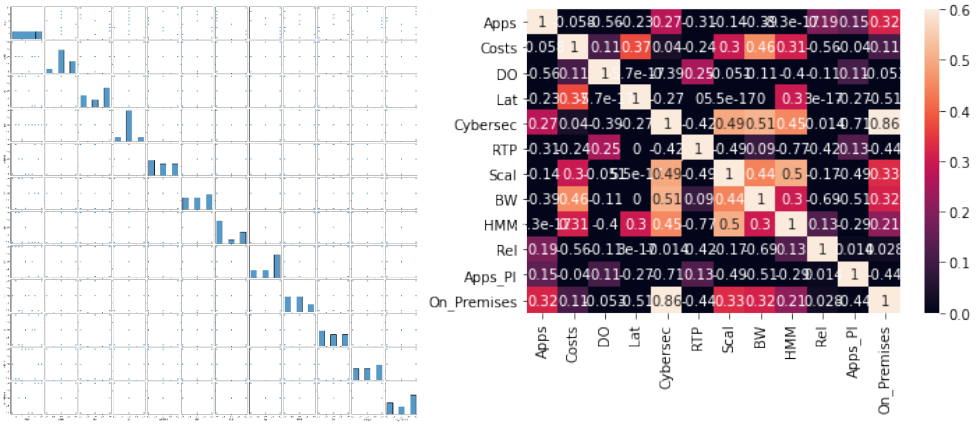

Figure 1: Correlation and Heatmap for on_Premises

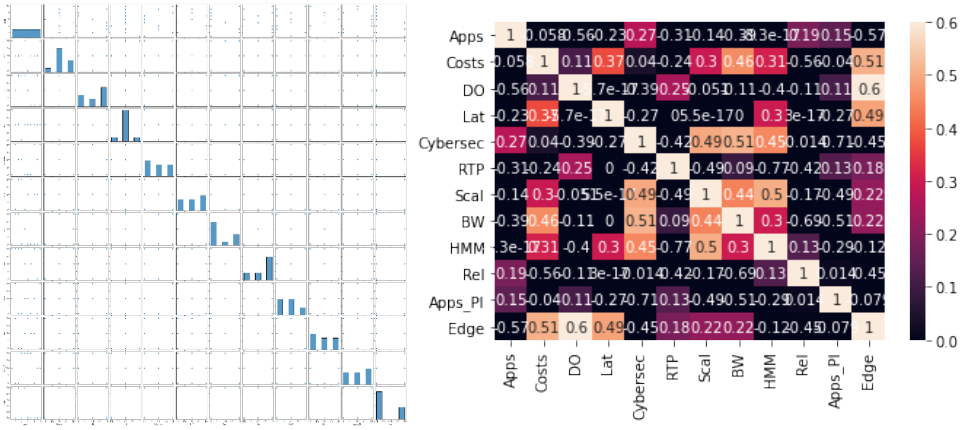

Figure 2: Correlation and Heatmap for Edge

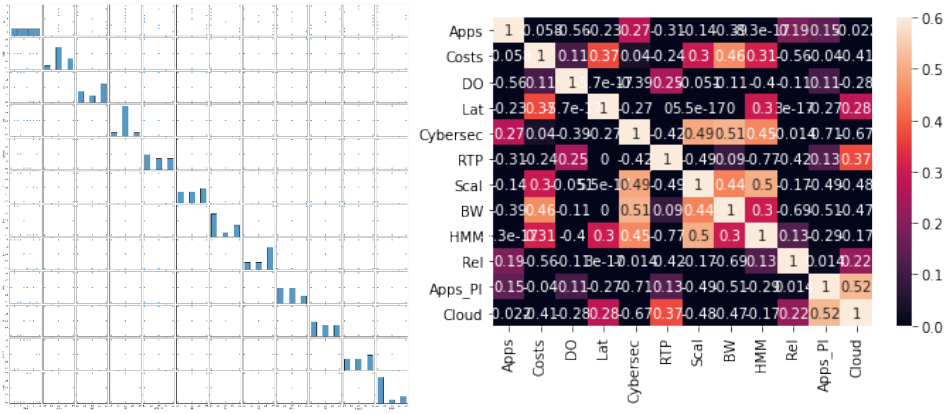

Figure 3: Correlation and Heatmap for Cloud

In this study, the Decision Tree algorithm was used to establish the relationships between the Applications (1-10), the criteria (see legend in Table 4.), and the three different computing environments, i.e., Cloud, Edge, and On_Premises. Training and testing data sets were split into a 75/25 ratio with the testing data sets being arbitrarily chosen, and in this case, they are Apps 3, 9, and 5. However, noticed that there was not much difference when considered at 60/40 training to test split. The goal was to predict the appropriate computing environment given the criteria established for the applications. The numbers for the criteria were generated randomly to recognize the disparity and differences in real world computing environments for different organizations depending on their size, resources, requirements, and Mission. The heatmap and correlation plots were selected to focus on the correlation between the parameters of interest. For the focus on cybersecurity, it is interesting to note that cybersecurity for performance-intensive apps for edge and cloud computing were

relatively close in correlation with a value of ~ 0.710. For the on-premises security environment, the correlation to the performance-intensive (including databases) apps was higher, to be expected, than the others with a value of ~0.860. There was an interesting observation with all three computing environments, wherein the RTP and HMM pairing yielded a value of ~ -0.77. Further analysis is required.

6. Conclusion

In this study, the focus was on addressing some of the major challenges associated with deploying applications within heterogeneous environments, including cloud, edge, and on-premises. A way to formally prioritize data and applications in terms of Mission importance was established. The efforts emphasized the relationship between Mission Critical applications, relevant data and storage locations, and the identification of the "right" compute resource for which an application is best suited. The applications, focus areas, and storage locations are all defined anonymously in this work and left to the determination of the organization. This work sought to prioritize data from top to bottom at different levels, define the location of the data in terms of importance—a continuing problem within many organizations that still do not know where a lot of their mission critical data is located. Knowledge of the most mission critical data to the least critical data allows organizations to more intelligently define and implement a budget for resources such as storage, as well as gain access to critical data when needed.

The correlation analysis was selected for this study to emphasize the similarity of variables of interest (especially as each variable relates to cybersecurity) when examining the criteria for defining the compute resource of choice. The correlations (positive and negative) are essential in reflecting how variables are trending in the same or opposite directions, respectively. The correlations were especially helpful in classifying applications according to their relevance to Mission. It is important to note that there will surely be different levels of the applications and data within the classes. For example, applications E3, E7, and E8 are each classified as high, but surely there is a hierarchy to the levels that can still easily be resolved through an additional application of the process until each application is addressed according to its Mission critical status from top to bottom. Future work will focus on defining a pipeline that allows the automation of the entire process, as well as, examining other machine learning algorithms, including an ensemble method to see how results compare between different algorithms.

References

Arora, S., 2022. *Edge Computing vs Cloud Computing: Key Differences to Know.* [Online]
Available at: https://www.simplilearn.com/edge-computing-vs-cloud-computing-article
[Accessed 19 01 2023].

Duggal, N., 2022. *Advantages and Disadvantages of Cloud Computing.* [Online]
Available at: https://www.simplilearn.com/advantages-and-disadvantages-of-cloud-computing-article
[Accessed 19 01 2023].

English, J., 2022. *What is Network Virtualization? Everything You Need to Know.* [Online]
Available at: https://www.techtarget.com/searchnetworking/What-is-network-virtualization-Everything-you-need-to-know#:~:text=External%20virtualization%20uses%20switches%2C%20adapters,without%20using%20an%20external%20network.
[Accessed 19 01 2023].

Kemper, F., 2021. *On-Premises vs Cloud: Advantages and Disasdvantages.* [Online]
Available at: https://www.empowersuite.com/en/blog/on-premise-vs-cloud
[Accessed 19 01 2023].

Noor, A. et al., 2019. *A Framework for Monitoring Microservice-Oriented Cloud Applications in Heterogeneous Virtualization Environments.* Milan, IEEE.

Ray, M., 2022. *Edge Computing: The Advantages and Disadvantages.* [Online]
Available at: https://innovatorscentral.ca/edge-computing-advantages-and-disadvantages/
[Accessed 12 01 2023].

Srivastava, P. & Khan, R., 2023. A Review Paper on Cloud Computing. *International Journals of Advanced Research in Computer Science and Software Engineering,* 8(6), pp. 17-20.

Varde, A., Bhonge, N., Duffield, M. & Morris, S. M., 2021. *Best Practices for Developing ClioSoft SOS on AWS,* Seattle: Amazon .