A Data-Centric, Defense-in-Depth Approach for Securing the Internet of Things

By Mangaya Sivagnanam – ISSA member, Minnesota Chapter

Big data drives significant benefits in understanding end user needs and proliferating new innovative businesses building new products and services. The data-centric approach discussed in this article ensures the right people get access to the right information, shared at the right time in the right channel.

Abstract

Big data drives significant benefits in understanding end user needs and proliferating new innovative businesses building new products and services to meet customers' expectations. The security and privacy challenges escalate because these devices tend to collect and retain so much data. Data mining can create far more attack patterns than most people can anticipate. The data-centric approach discussed here ensures the right people get access to the right information, shared at the right time in the right channel.

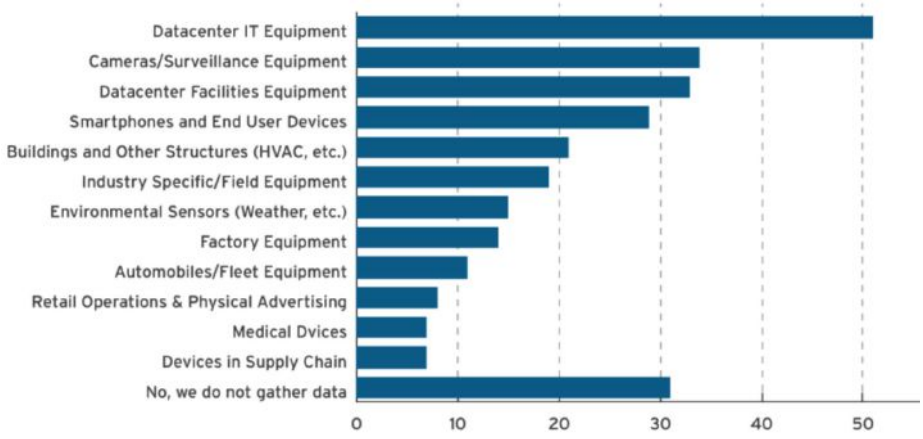

Figure 1 – A Snapshot of data collected in smart connected buildings
(Source: 451 Research, Voice of Enterprise Internet of Things)

Smart-connected buildings represent a convergence of latest cutting-edge technologies including self-awareness, predictive maintenance, convergent networks, wireless retrofits, and biometrics. These buildings are smartly equipped and connected with all flavors of Internet-connected cyber-physical systems (CPS, popularly known as the Internet of Things—IoT) ranging from small sensors to large industrial control systems.

The IoT's ability to collect, analyze, retain, and transform data drives much of the value of IoT devices and services and functionality provided. However, the current processes followed in data collection, retention, and processing have disadvantages and privacy issues.

Also, currently there are no unified industry or manufacturer-specific standards, regulations, or processes enforced on smart-connected buildings and building automation systems (BAS). As the big data environment proliferates, most organizations tend to focus only on the potential benefits and fall behind in establishing and maintaining the data appropriately.

The volume of data collected and retained in various endpoints in smart-connected buildings is immense. Figure 1 shows a snapshot of the data gathered in the building network for a medium-size organization. The research breaks data collected into three broad categories: machine sensing (71.5 percent data collected from systems, CPS), biological sensing

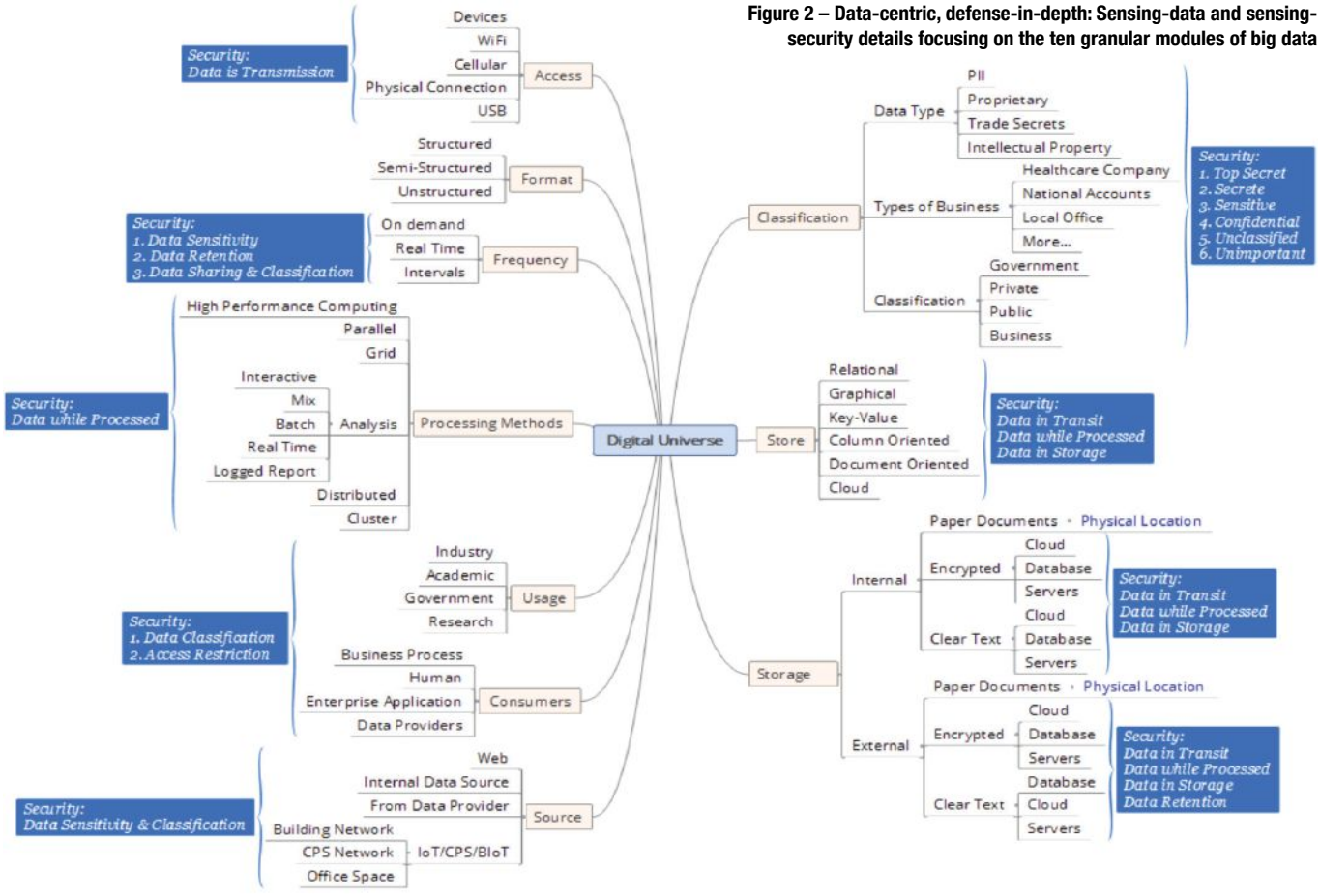

Figure 2 – Data-centric, defense-in-depth: Sensing-data and sensing-security details focusing on the ten granular modules of big data

(8.5 percent data collected from humans and animals), and environmental sensing (20 percent data collected from the environment) [2].

Handling big data comes with significant responsibilities. The following are critical challenges with big data in the organization:

- Abundant and indiscriminate data collection from a wide range of devices
- Unexpected uses of collected data, especially without any consent
- Unintended data breach risk with substantial consequences
- Outdated approaches – security
- Insufficient governance

The "data-centric, defense-in-depth" approach improves big data lifecycle security. Figure 2 shows the graphical representation of sensing-data and sensing-security details, focusing on the ten granular modules of big data.

The approach classifies big data into ten granular modules: Data access, data format, data frequency, data processing methods, data usage, data consumers, data source, data classification, data store, and data storage. These classifications facilitate implementing a data-centric security approach with three primary security controls: Preventive, detective, and administrative.

Preventive

The preventive security control is the process of securing the data with controls such as encryption of data while the data is in transit, in process, and at rest (stored). It also enforces implementation of identity and access management on the data, based on its classification level.

Detective

The detective security controls look for anomalous behavior in the CPS device and its network. Organizations need to place Hadoop services, data activity, and monitoring systems on respective places in the building network, based on the sensitivity classification of the data.

Administrative

The administrative security controls implement security tools to benefit people, process, or technology: Sensitive data classification, user privilege management, configuration management, encryption, key management, and more. Administrative controls also help define separation of duties, authentication, and authorization of users, databases, servers, and applications.

Module usage in a data-centric security approach

- "Classification," "Usage," and "Consumers" define the sensitivity of the data based on the business needs and organizational risk metrics.

- "Source," "Storage," "Store," and "Processing Methods" define security controls for data in various states (transit, process, rest). They help build security monitoring, auditing, alerting, and reporting.
- "Source," "Access," "Frequency," "Classification," and "Format" help the organization define appropriate administrative security controls.

The model recommends an organization defines a risk rating for all 10 data modules, based on the organizational risk strategy. The ratings help the organization strengthen security controls as needed in the right place.

SAMPLE DATA	CONFIDENTIALITY	INTEGRITY	AVAILABILITY	PRIORITIZATION : HIGH (1) – LOW (9)
Username	Medium (some cases email address is username)	Medium	Medium	5
Password	Critical			1
SSN	Critical			1
Phone number	Low	Low	Medium	8
PCI Info	Critical			1
Corp. address	Low (might already be available on the Internet			9
Corp. info	Low	Medium	Medium	8
User Hobbies	Low (available in social media network)			9
Server Room	Low (physical proximity)			7

Table 1 – Big data creation and discovery—prioritization of data

Secure life cycle for big data

The life cycle of big data has six main stages: creation and discovery, access and data flow, process, share, store, and destroy. The data-centric, defense-in-depth approach federates the data movement and helps implement proper preventive, detective, and administrative controls in every stage of the big data life cycle as discussed here.

1. Creation and discovery

Creation and discovery is the first and foremost stage in the life cycle of big data. Effective data classification will benefit the organization by implementing proper security controls in the big data environment. However, the diversity of data sources, data flows, and data formats in the high volume environment creates security risks and implementation challenges.

The key challenges in this stage are:

- Identifying all end points in the network that contribute the data [4]
- Identifying intellectual property and determining the value and business impact of each data in the big data cluster
- Defining data provenance

Lack of awareness of the sensitivity of the data will lead to exposure of significant risks. All data gathered from various end points (systems and building network) need to be identified and formatted appropriately, either as structured or unstructured format.

Structured data is easy to manage and maintain. On the other hand, unstructured data needs to be correctly formatted to match the security requirements so that it could be easily secured.

For example, it is easier for the organization to discover structured sensitive data like relational data created in databases, JavaScript object notation (JSON), comma-separated values (CSV), and XML-formatted files and data. With unstructured data, it is difficult to discover and locate the sensitive data in the data clusters, so it is crucial to define the structure in the data [4]. Table 1 discusses a sample exercise of how personal user data can be classified and prioritized based on its sensitivity.

2. Access and data flow

After discovering and classifying the data by its structure and confidentiality, we need to identify its flow and linkage to the system. All data are meant to be accessed by various systems and individuals inside and outside the corporate network. It is essential to recognize all possible data boundaries/zones distinguishing the end-to-end data flow in the network. These virtual network boundaries will facilitate constructing a secure big data environment.

The security challenges in this stage are:

- Implementing security in distributed programming frameworks
- Implementing granular access controls (sensitive data vs. user role)
- Defining security controls for non-relational data sources
- Identifying end-to-end data flow

Producing a data-centric threat analysis will help us to identify the data flow, linkage, and data necessity across distributed programming frameworks. Also, based on the prioritization matrix from the previous phase, this will help govern the right people getting access to the data. Identity and access management with granular, role-based access controls and access control list, for active directory, systems configuration files, application accounts, and system root determines that the *right people access the right information*.

The data governance is no good if the data transferred through the network is insecure. Cleartext data might be transferred in the network. Governance must enforce the organization to use correct data access channels like HTTPS/TLS to improve network security and privacy.

Understanding the end-to-end data flow in the big data environment aids in defining and understanding the egress and ingress methods in order to strengthen network boundaries.

Centrally managing data access policies will leverage attribute-based access controls and can protect data based on tags and data lineage (data origins, movements, characteristics, and quality) [14].

3. Process

Data mining and analytics techniques process the abundance of varied data to obtain meaningful information and intelligence to meet business needs. The organization needs to pay extra attention to securing the data while processing. A user can access and process the data either inside or outside the building network via the Internet and various cloud services offered by third-party service providers.

Security challenges while data processing are:

- Implementing scalable privacy and security during data mining and data analytics
- Implementing granular data audits

The availability of data is critical in this phase to perform timely analytics on the data to satisfy the business needs. The critical success of data processing determines that the *right people* get access at the *right time* in the correct channel. In addition to security controls recommended in the big data access phase, user entitlement and data metering can be used to improve security while processing the data.

The entitlement granted to a particular user or data can restrict the specific user account on the target system to perform specific analytics or functions on big data cluster environment [7].

The metering of data can regulate the total amount of data that can be utilized concurrently in one analytics algorithm or process flow [8]. Also, defining a time-to-live state on each data processing analytic in the network provides additional protection by regulating misuse of data and its computation by adversaries.

4. Share

Collected data might be shared with all employees in the building network and external entity. Implementing logical internal/external security access controls with appropriate secured data-access channels improves confidentiality, integrity, and availability.

The security challenges while sharing data are:

- Implementing granular data audits
- Implementing reactive security to secure the integrity of data

In addition to security controls recommended in the previous stages of big data life cycle, packet collections, folder/file encryption, digital signature, hardware security module (HSM), application level data encryption, physical security and volume control, audit logging, and data hashing are preferred in this phase.

Encryption, digital signature, and data hashing will protect the integrity of the data while sharing confidential information in the network or cloud.

HSM and full-disk encryption are the best defense options to consider to maintain hardware trust and integrity of data while sharing sensitive information. Encryption at the hardware/folder/file/data level can offer granularity in security and audit logging capabilities.

5. Storage

Organizations tend to store data for an extended period. The stored data are processed and analyzed to serve organizational needs better. This intelligence makes the building even smarter. So the organization must place proper security access controls in data backups. The data at rest should also be encrypted to maintain the confidentiality of data.

The security challenges while storage data are:

- Implementing secure data storage and transactional data logs and files

To fortify data protection and privacy, the organization should implement data storage, loss, and data retention policies.

Data storage – all sensitive data should be hashed, masked, or encrypted, depending on the sensitivity while storing.

Data retention – depending on the industry needs periodic backups and automatic or manual data disposal features

VIRTUAL DATA BOUNDARIES	POSSIBLE THREATS	POSSIBLE SECURITY CONTROLS
Untrusted zone (customer network)	External Internet-based threats; malware in user applications (laptops, mobile, tablets…)	1. **Identity and access management:** User entitlement, data metering, ingress/egress methods 2. **Network security:** Authentication, LDAP, RBAC, Kerberos, auditing, activity monitoring 3. **Secure data access channels:** HTTPS, TLS 4. **Secure data sharing channel:** HTTPS, TLS, privileged user, encrypted, time-to-live of data 5. **Role-based access control – RBAC** 6. **Encryption:** Cryptographic protection, field-level/column-level encryption, packet-collection encryption, data masking, data hashing 7. **Integrity protection:** Hashing, digital signature 8. **Physical security:** Locks, fireproof/water safe... 9. **Hardware Security Module** 10. **Secure data disposal policy** 11. **Remove user/system access rights policy**
Trusted zone (enterprise network)	Malware, virus, Trojan…in network, applications, and workstations, email phishing attacks, malicious downloads, active directory exploits, insider threats, identity theft	
Trusted zone (Sensitive data storage and processing zone)	Exploit of cryptographic weakness, LDAP/Kerberos weakness exploits, Windows/Linux-based attacks, server exploits, database/injections attacks (OWASP attacks)	
Trusted zone (big data clusters zone)	Location/sensitivity-based data exploits in unstructured data, server exploits, cyber simulations in ingress points, cyber simulation/malfunctioning of data analytics exploits	
Hardware trust zone	Public-private key exploits, key-management server exploits	
Workstation/ application zone	External Internet-based threats, malware in user applications (laptops, mobile, tablets…), malware, virus, Trojan… in network, applications, workstations	

Table 2 – Data-centric threat model and big data security framework

should be in place to improve data protection. All server rooms and data storage servers should have proper physical security.

6. Disposal

Data disposal is the most crucial stage in the life cycle of big data. Data in the wrong hands may be catastrophic. Organization-level security policies to implement secure data disposal methods and removal of access rights on employee/user exit interviews should be in place to ensure the data is available only to authorized users.

Data-centric threat model and big data security framework

A data-centric threat model focusing on end-to-end data flow while pinpointing confidentiality, integrity, availability, threats in the network boundaries, value of the data, and business impact will help the organization identify appropriate security controls in the big data life cycle. The model also helps the organization quantify the risks and threats to customer data and the business.

Articulating the results from data discovery and classification discussed above in the big data life cycle and recognizing industry-known threats in each data zone will facilitate organizations to implement appropriate security controls. Table 2 categorizes possible threats in each virtual data boundary.

Big data and smart-connected buildings – a snapshot

Smart-connected buildings are smartly equipped and connected with all flavors of Internet-connected cyber-physical systems ranging from small sensors to large industrial control systems. The proliferation of connected products and applications in the network will increase the data collected to support building analytics, which will increase the attack surface and potential data theft in the organization.

The example discussed here elaborates how the data-centric defense-in-depth approach and data-centric threat model will have a positive impact and improve security in big data management.

Studies show that on average data collected by smart buildings is roughly 20 zettabytes of data in 2018 through the range of sensors and smart-connected devices[12].

The six lifecycle stages explained earlier will facilitate effective data discovery and classification based on sensitivity and associated threats, helping us define adequate structure and format to the data. The proper classification will help manage the complex forest of data and help outline data boundaries.

The different data boundaries considered in this scenario are un-trusted, trusted/corporate, sensitive-data trusted, big data cluster, hardware trust, and physical. The classifications assist the business to restrict sensitive data access in the un-trusted boundaries. At the same time it will help the security analyst to implement a proper identity and access management system, network security, and RBAC in secure data-access channels. The process will help limit the digital footprint of sensitive data on the Internet and local network, which makes it easier to monitor and control data movement.

In this setup, sensitive data is limited to "sensitive-data trusted zone" and "big data cluster" to enforce additional safety to improve integrity and confidentiality of the data. The method shows an apparent transformation of substantial, unmanageable forest data to manageable big data, which Hadoop can effortlessly handle. Furthermore, data movement is limited to appropriate virtual network boundaries based on sensitivity. This approach helps us sense the granular information and security requirements for the data, resulting in affordable and scalable security.

Figure 3 is a snapshot of data transformation in a connected-building network with the help of this model. It elaborates some of the possible physical and logical security controls that can be implemented at every stage of the big data life

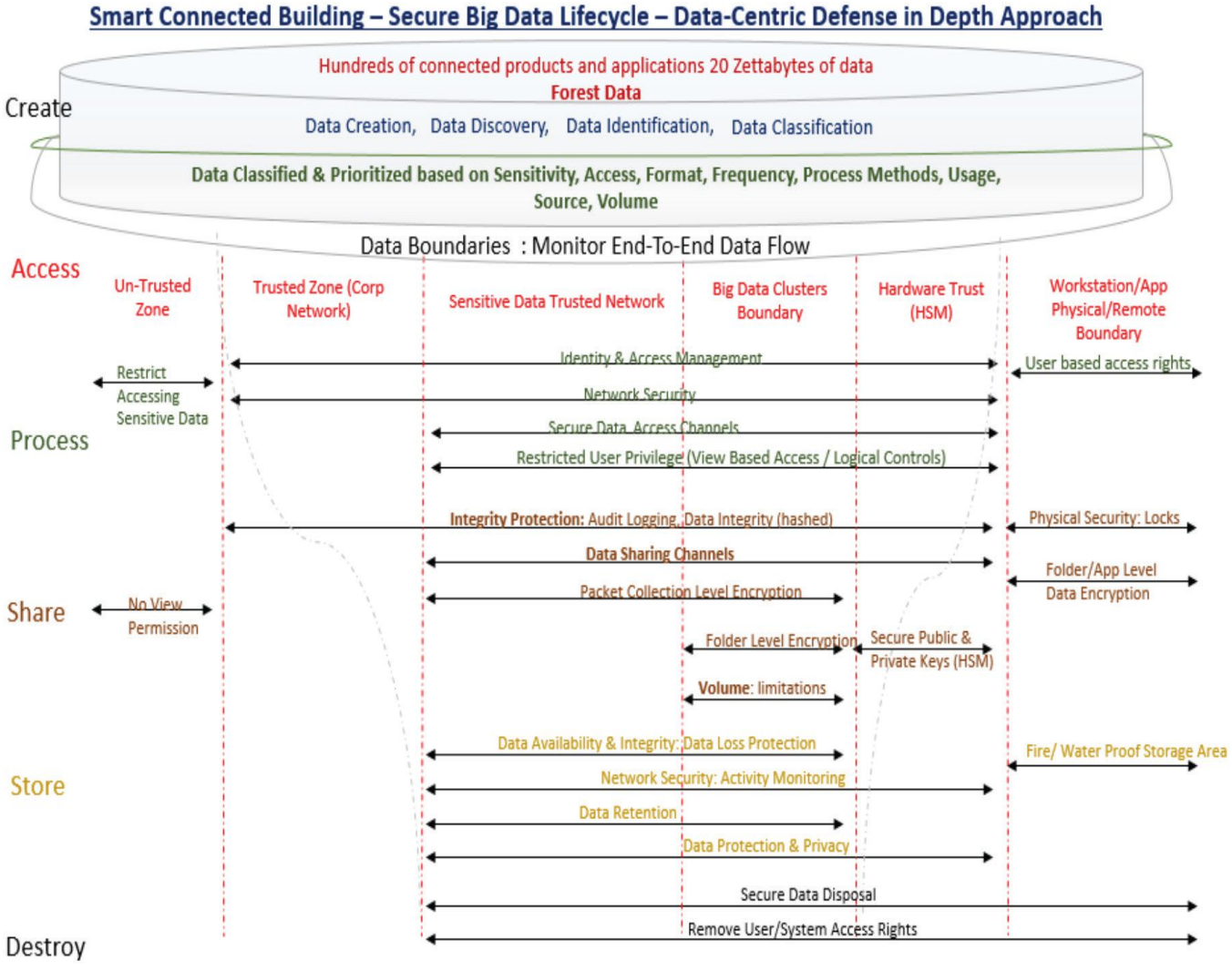

Figure 3 – Big data and smart-connected building snapshot [data-centric defense-in-depth approach]

cycle in various virtual network boundaries. The selection of the security controls depends on industry needs, volume, and diversity of the data available in the network [13].

Conclusion

Handling big data comes with significant responsibilities; the more massive the concentration of sensitive personal data, the more attractive to criminals. The risk of compromising sensitive information increases as the volume and sensitivity of the data grows. Protecting the data at every stage of its life cycle is essential. Creating a threat model focusing on the data in each stage (create, access, process, share, store, destroy) will benefit the organization, quantifying the risks and threats in customer data and the business. The data-centric approach discussed here ensures that the *right people* get access to the *right information*. Also, make sure the right information is shared at the right time and place in the right channel, thus implementing preventive, detective, and reactive security in the big data environment that will promote the successful functioning of the businesses.

References

1. Oracle, "Securing the Big Data Life Cycle," Oracle (accessed on Apr 24, 2018) – http://files.technologyreview.com/white-papers/Oracle-Securing-the-Big-Data-Life-Cycle.pdf.
2. Essery, M., "Today 65% of Enterprises Already Using Internet of Things; Business Value Found in Optimizing Operations and Reducing Risk," 451 Research, 29 June 2016 (accessed on Apr 24, 2018) – https://451research.com/blog/419-today-65-of-enterprises-already-using-internet-of-things-business-value-found-in-optimizing-operations-and-reducing-risk.
3. IEC, "IEC 62351: Power Systems Management and Associated Information Exchange - Data and Communications Security - Part 8: Role-based Access Control," International Electrotechnical Commission – https://webstore.iec.ch/preview/info_iec62351-8%7Bed1.0%7Den.pdf.
4. Alshboul, Y. et al, "Big Data Life Cycle: Threats and Security Model," Emergent Research Forum papers – https://pdfs.semanticscholar.org/9ef4/fa7c505b92a5a0a9621fb5646b9d70739d27.pdf.

5. Gaddam, A., "Securing Your Big Data Environment," Black Hat USA 2015 – https://www.blackhat.com/docs/us-15/materials/us-15-Gaddam-Securing-Your-Big-Data-Environment.pdf.
6. Hash, J. et al, "NIST Special Publication 800-65: Integrating IT Security into the Capital Planning and Investment Control Process (document accessed on Apr 24, 2018) – http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-65.pdf.
7. IBM, "Understanding the Data Entitlement Object Model, IBM Knowledge Center – https://www.ibm.com/support/knowledgecenter/en/SSWSR9_11.6.0/com.ibm.mdmhs.dev.platform.doc/concepts/c_Data_Entitlement_Object_Model.html.
8. Oracle, "Utilities Meter Data Management," Oracle – https://www.oracle.com/industries/utilities/products/meter-data-management/index.html.
9. NIST, "Welcome to NIST Big Data Public Working Group," NIST Big Data Public Working Group – https://bigdatawg.nist.gov/.
10. NIST, "Big Data Interoperability Framework: Volume 2, Big Data Taxonomies," NIST Big Data Public Working Group – https://bigdatawg.nist.gov/_uploadfiles/M0636_v1_3062599334.docx.
11. Hadoop, "MapReduce Tutorial," Hadoop – https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
12. Statista, "Volume of data Collected by Smart Buildings Worldwide from 2010 to 2020 (in zetabytes)," The Statistics Portal – https://www.statista.com/statistics/631151/worldwide-data-collected-by-smart-buildings/.
13. Dell/EMC, "Security and Compliance for Scale-Out Hadoop Data Lakes," EMC – https://www.emc.com/collateral/white-paper/h13354-wp-security-compliance-scale-out-hadoop-data-lakes.pdf.
14. Michelle Knight, "Data Lineage Demystified: The What, Why, and How," Dataversity, April 20, 2017 – http://www.dataversity.net/data-lineage-demystified/.

About the Author

Mangaya Sivagnanam is currently Principal Cybersecurity Systems Architect at Ingersoll Rand with 17 years experience in web, enterprise, and embedded applications framework design, analysis, development, and deployments on client/server applications and commercial industrial control systems. She may be reached at s.mangaya@gmail.com.