Data Is the New Bullet

WHITE PAPER: AUGUST 2021

AUTHORS:
Matthew D. Gonzalez, Ph.D.
Jose Rodriguez IV

For organizations, what is the most valuable asset besides their human force, knowledge, and revenue? In our opinion, it is their data. Without data, organizations could not sell their services or products, stay competitive, make important decisions and generate profits.

Through the adoption of data and analytics strategies and best practices, data can help refine your strategy, place the right data in the hands of the appropriate decision makers, while ensuring it remains:

- organized,
- readily accessible,
- classified, and
- secured using leading edge technologies and industry best practices.

AUDIENCE

This white paper is intended for IT executive-level management personnel and IT SMEs working for government organizations as well as enterprise-level private-sector companies and nonprofits.

DATA IS THE NEW BULLET

A loaded gun, with ammunition, can save an individual or group of individuals from being killed. As well, a properly-implemented and securely-architected data management strategy can proactively shield an organization from being taken hostage by internet hackers who can compromise their data. The embarrassment and bad publicity that can potentially lead to losing an organization's loyal base.

DATA AND ANALYTICS: TOWARD REFINED STRATEGY AND DECISION MAKING

> *"91 percent of organizations have not yet reached a 'transformational' level of maturity in data and analytics[1]."*
>
> \- GARTNER -

Your first, or next, 100 days toward strategy development and decision making should largely incorporate your data and analytics (D&A) strategy. **This starts with baselining where you are today regarding authoritative versus non authoritative data sources, identifying where you want decisions to come from, then closing the gap by:**

- Involving your data team through data analytics working groups (DAWGs)
- Adopting data governance models, processes, and procedures
- Integrating technology (e.g., bots, data visualization tools, intelligence, etc.),
- Assessing data maturity
- Measurement through D&A Scorecards *(sample in Appendix B)*

Data is the means to the end, while analytics helps shape the end decision points. Thus, one must help their employees "speak data and analytics" through active involvement mixed with technological solutions that add value toward refined decision making.

DATA LAKES AND DATA WAREHOUSES

A recent article on FedTechMagazine defines data lakes and data warehouses as follows:

> "Data lakes are different in structure and function from data warehouses. Here are some of the key differences between the two.
>
> A **data lake**, as Oracle notes in a blog post, is "a place to store your structured and unstructured data, as well as a method for organizing large volumes of highly diverse data from diverse sources." Data lakes often can ingest data very quickly and then "prepare it later, on the fly, as people access it."
>
> In contrast, a **data warehouse** "collects data from various sources, whether internal or external, and optimizes the data for retrieval for business purposes," Oracle notes. Data in data warehouses is often structured and usually comes from relational databases, but it can be unstructured too. Primarily, Oracle's blog adds, "the data warehouse is designed to gather business insights and allows businesses to integrate their data, manage it, and analyze it at many levels[2]."

Therefore, securely storing and classifying data either on-premise at the organization's data centers or with a Cloud Service Provider (CSP) such as Microsoft Azure or Amazon Web Services (AWS) is a must so the organization can securely access these data lakes and data warehouses and data can be accessible employing data protection best practices.

The definitions of structured, semi-structured and unstructured data as defined by Microsoft more technically in nature are described in "Appendix A" of this article.

MICROSOFT AZURE AND AMAZON WEB SERVICES (AWS) DATA LAKES

Renting aka Pay as you Go (PAYG) computing, storage, networking and security services from CSPs is becoming more prevalent and more widely accepted by government agencies due mainly to the construction of cloud government data centers that adhere to higher levels of security in depth controls.

Microsoft offers four secured government data centers strategically located in Virginia, Texas, Iowa and Arizona with many high availability, resiliency, and redundancy features, most of them for a consumption-based price. On the other hand, AWS provides similar cloud services and tool sets for West and East Coast government customers.

Microsoft Azure advertises the following Data Lake attributes on their website as follows:

> "Data Lake Analytics is an on-demand analytics job service to power intelligent action. Process big data jobs in seconds with Azure Data Lake Analytics. There is no infrastructure to worry about because there are no servers, virtual machines, or clusters to wait for, manage, or tune. Instantly scale the processing power, measured in Azure Data Lake Analytics Units (AU), from one to thousands for each job. You only pay for the processing that you use per job.
>
> Process petabytes of data for diverse workload categories such as querying, Extract, Transform, and Load (ETL), analytics, machine learning, machine translation, image processing, and sentiment analysis.[3]"

On the other hand, Microsoft archrival AWS' Data Lake selling point is:

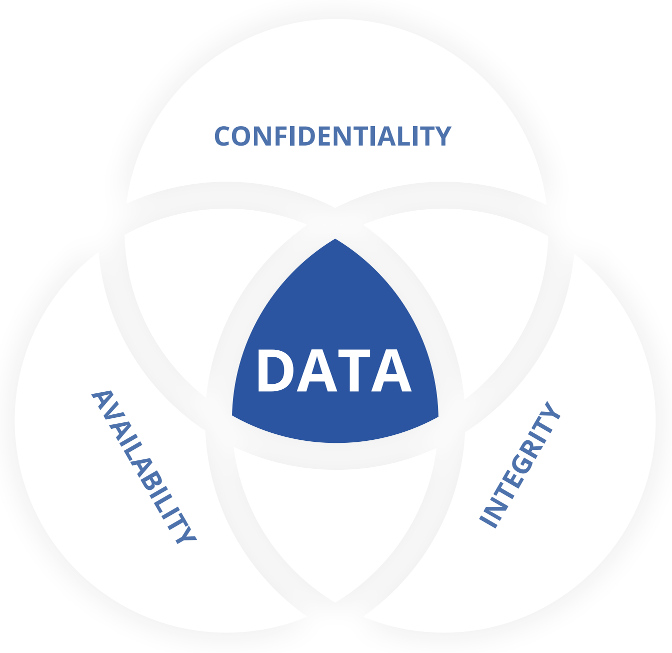
> "Many Amazon Web Services (AWS) customers require a data storage and analytics solution that offers more agility and flexibility than traditional data management systems. A data lake is a new and increasingly popular way to store and analyze data because it allows companies to manage multiple data types from a wide variety of sources, and store this data, structured and unstructured, in a centralized repository.
>
> "AWS offers a data lake solution that automatically configures the core AWS services necessary to

> **easily tag, search, share, transform, analyze, and govern specific subsets of data across a company or with other external users. The solution deploys a console that users can access to search and browse available datasets for their business needs. The solution also includes a federated template that allows you to launch a version of the solution that is ready to integrate with Microsoft Active Directory.[4]"**

CIA TRIAD

Notwithstanding the importance of utilizing Data Lakes and Data Warehouses, one cybersecurity concept that has been widely accepted and implemented across both the public and private sectors is the Confidentiality, Integrity, and Availability (CIA) triad. The idea behind the CIA triad is that in order to pro-actively (vs. passively) protect the organization's data assets, a three-dimensional approach consisting of CIA is required to put in place, maintained and monitored by the organization's stakeholders.

First, the data *availability* aspect plays a pivotal role because it does not matter whether the data is secure, it needs to be *available* to only "authorized" data consumers, whether internal or external to the organization, to use, meet their needs and solve their problems. For example, internally to the organization data needs to be *available* to produce dashboards for data visualization and decision-making reports. Externally, data consumers now want millisecond data *availability* over the internet to file their taxes or register on a federal agency website. If an internet commerce website loads too slow and consumers have to wait several seconds for content to load, they may opt to access a different online merchant to find and buy what they are looking for.



Second, data *integrity* is key for data consumers and producers to successfully make use of their data. In layman's terms, data *integrity* is having data that is not erroneous, duplicated, and/or corrupted. For instance, in the healthcare industry, maintaining accurate medical records about a patient can significantly impact the positive or negative outcomes for the patient's speedy recovery.

Third, data *confidentiality* requires the implementation of logical and physical layers of cybersecurity controls that will be explained next under the "Defense in Depth" section in this white paper, but holistically speaking *confidentiality* is comprised of a wide array of cybersecurity mechanisms that work in tandem to ensure data assets are ironclad protected.

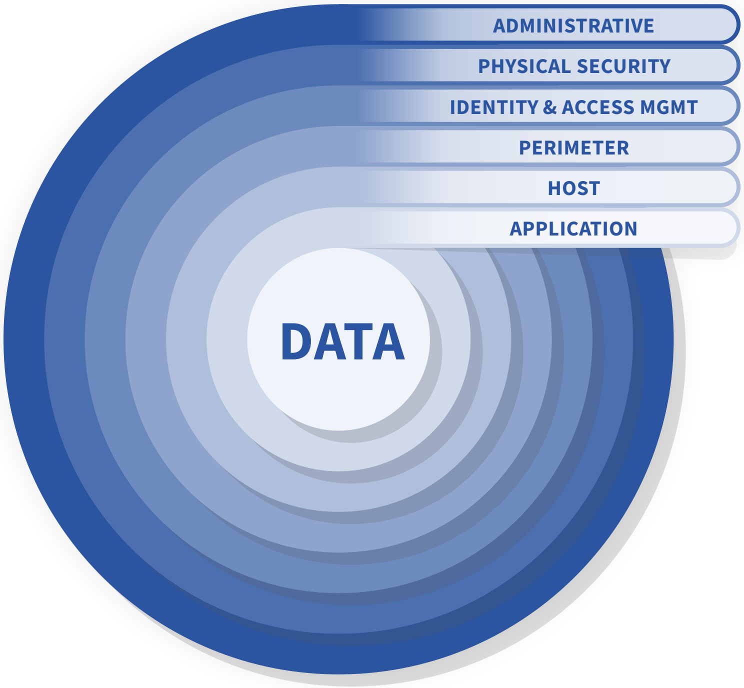
DEFENSE IN DEPTH

Defense in Depth relates to the fact that in order to truly protect data assets, a multi-layered cybersecurity approach needs to be implemented. A single or pick and choose layer cybersecurity architecture is weak because it leaves empty holes that can potentially be exploited by the cyber criminals.

Defense in Depth is defined by the International Information System Security Certification Consortium, or ISC2 Lexicon as follows:

> **"The coordinated use of multiple safeguards – which can be administrative, physical, and/or technical – to protect the confidentiality, integrity and availability of information assets and systems. A key principle within Defense in Depth is that there should be compensating controls in place to defend a system if one countermeasure fails.[5]"**

Data whether Personally Identifiable Information (PII), Protected Health Information (PHI), finance, or intellectual property, which is what hackers are after and they can buy or sell in the dark web, resides at the core of the Defense in Depth model and it is protected by six outer security layers. Please note that various vendors and organizations define the Defense in Depth model based on their understanding and perspectives, but almost all across the board agree that data lives at the center of the model surrounded by similar but not equal variations of the outer security layers.



Using this framework, bullet-proof data protection is only achievable if collaboration exists among the different departments overseeing the Administrative, Physical Security, Identity and Access Management (I&AM), Perimeter Security, Host Best Practices, the Application and Data Security aspects of this spectrum. Implementing Physical Security controls such as cameras, locks and badge readers do no good if Administrative controls to train staff to distinguish between legitimate or non-legitimate email attachment downloads or to practice cybersecurity hygiene by scanning external USB drives, if these are permitted within the organization.

USE CASE:

US AIR FORCE PAVING THE WAY TOWARDS MORE RESILIENT DATA PROTECTION CAPABILITIES

The US Air Force has been pioneering, leading, and deploying data management best practices to enhance their Command and Control (C2) capabilities and empower their airmen as demonstrated when they launched their VAULT platform:

> "The Air Force Chief Data Office announced today additional capabilities for the Visible, Accessible, Understandable, Linked and Trusted Data Platform. The VAULT Platform is designed to provide Airmen with cyber secure, cloud-based tools to connect, find, share and learn from Air Force data to improve readiness and mission success."

> "The VAULT Platform now has a set of tools to support a full lifecycle of data exploitation activities. It is now possible to ingest data, manage storage, manage metadata, manipulate, cleanse and experiment with data and visualize analytics results. The data sets and the applications have access controls that today are group and role based. The visualization service supports multiple user groups leveraging a common application platform.[6]"

CONCLUSION

Whether your organization is in the public or private sector, a small IT shop or a conglomerate, or if your data centers are hosted on your premises or in the cloud, incorporating the aforementioned data and analytics strategy recommendations, cybersecurity guidance, and implementation best practices can strengthen your arsenal of bullets, provide the accurate calibers to protect your data, and shield it from the evil doers on the dark web.

APPENDIX A:

MICROSOFT DATA TYPE DEFINITIONS[7]

Structured: Sometimes referred to as *relational data*, is data that adheres to a strict schema, so all the data has the same fields or properties. The shared schema allows this type of data to be easily searched with query languages such as SQL (Structured Query Language). Structured data is straightforward in that it's easy to enter, query, and analyze. All of the data follows the same format.

Semi-structured: is *less* organized than structured data, and is not stored in a relational format, as the fields do not neatly fit into tables, rows, and columns. Semi-structured data contains tags that make the organization and hierarchy of the data apparent - for example, key/value pairs. Semi-structured data is also referred to as non-relational or NoSQL data. The expression and structure of the data in this style is defined by a serialization language such as XLM, JSON, or YAML.

Unstructured: It is ambiguous. Unstructured data is often delivered in files, such as photos or videos. The video file itself may have an overall structure and come with semi-structured metadata, but the data that comprises the video itself is unstructured. Therefore, photos, videos, and other similar files are classified as unstructured data.

APPENDIX B:
D&A SCORECARD SAMPLE

D&A SCORECARD	
DOMAIN 1: CREATE THE D&A VISION AND STRATEGY	**DOMAIN 1: MATURITY LEVEL**
Forge the Vision	5
Design the Strategic Plan	2
Create the Functional Design	2
Implement the Strategy	1
DOMAIN SCORE	2.50
DOMAIN 2: MANAGE THE D&A FUNCTION	**DOMAIN 2: MATURITY LEVEL**
Prioritize Project Proposals	4
Manage Projects	3
Monitor Portfolio Health	3
DOMAIN SCORE	3.33
DOMAIN 3: ALIGN D&A TO BUSINESS OUTCOMES	**DOMAIN 3: MATURITY LEVEL**
Establish a KPI and Metrics Framework	3
Quantify the Value	2
Innovate the Business Model	3
DOMAIN SCORE	2.67

SOURCES

[1] https://www.gartner.com/en/newsroom/press-releases/2018-02-05-gartner-survey-shows-organizations-are-slow-to-advance-in-data-and-analytics

[2] https://fedtechmagazine.com/article/2019/01/data-lakes-what-they-are-and-how-they-can-benefit-feds-perfcon

[3] https://azure.microsoft.com/en-us/services/data-lake-analytics/?&ef_id=EAIaIQobChMIofyRt8Po8QIVkojICh0SXwXjEAAYASAAEgIriPD_BwE:G:s&OCID=AID2200277_SEM_EAIaIQobChMIofyRt8Po8QIVkojICh0SXwXjEAAYASAAEgIriPD_BwE:G:s&gclid=EAIaIQobChMIofyRt8Po8QIVkojICh0SXwXjEAAYASAAEgIriPD_BwE

[4] https://aws.amazon.com/solutions/implementations/data-lake-solution/#:~:text=AWS%20offers%20a%20data%20lake,or%20with%20other%20external%20users.

[5] https://www.isc2.org/-/media/479EDA1AA73F49268997E1ADC9FB0740.ashx

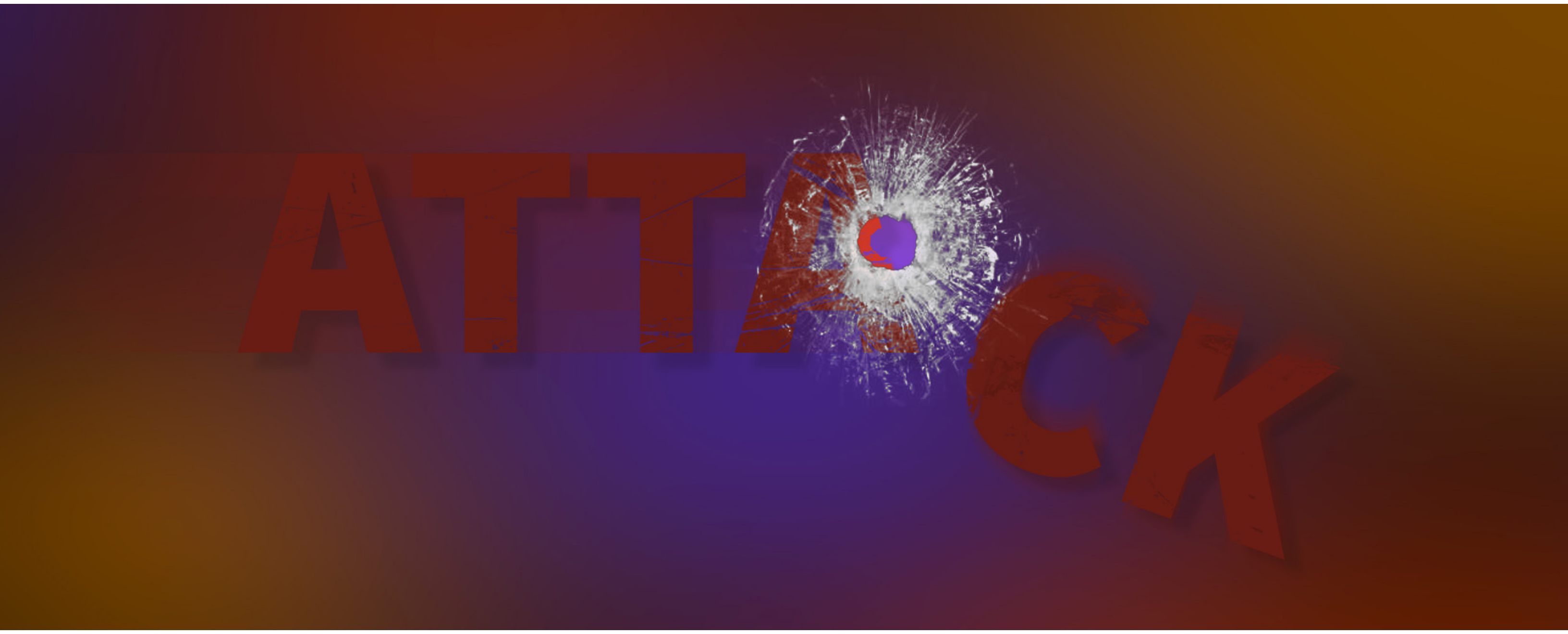
[6] https://www.af.mil/News/Article-Display/Article/1987254/chief-data-office-announces-capabilities-for-the-vault-data-platform/

[7] https://docs.microsoft.com/en-us/learn/modules/choose-storage-approach-in-azure/2-classify-data

ABOUT THE AUTHORS

Matthew D. Gonzalez, Ph.D. is an IT and Data SME. He has spent 25 years in IT working in various capacities in Fortune 500 companies, plus the DoD, from Developer, IT Architect, IT Project Manager, Chief Data Officer, and Associate Professor. His research in the areas of blockchain, and leadership in IT, have led him to publish and present at Harvard and the Pentagon.

Jose Rodriguez IV is Webhead's DevSecOps Officer. He has over 30 years of IT experience, including 20 years as a contractor working for federal organizations such as MHS, DISA, Army MEDCOM, DHMSM, USUHS, DHA, and USAF AETC. His technology focus areas specifically have been to Cybersecurity, IT architecture, cybersecurity and DevSecOps (DSO).



Web-Hed Technologies, Inc. (dba Webhead)
1710 N. Main Ave • San Antonio, Texas 78212
(210) 354-1661 • webheadtech.com