EBOOK

Cybersecurity in the Public Sector

Protecting the U.S. government with advanced analytics and AI

databricks

Contents

databricks

INTRODUCTION

The State of the Industry

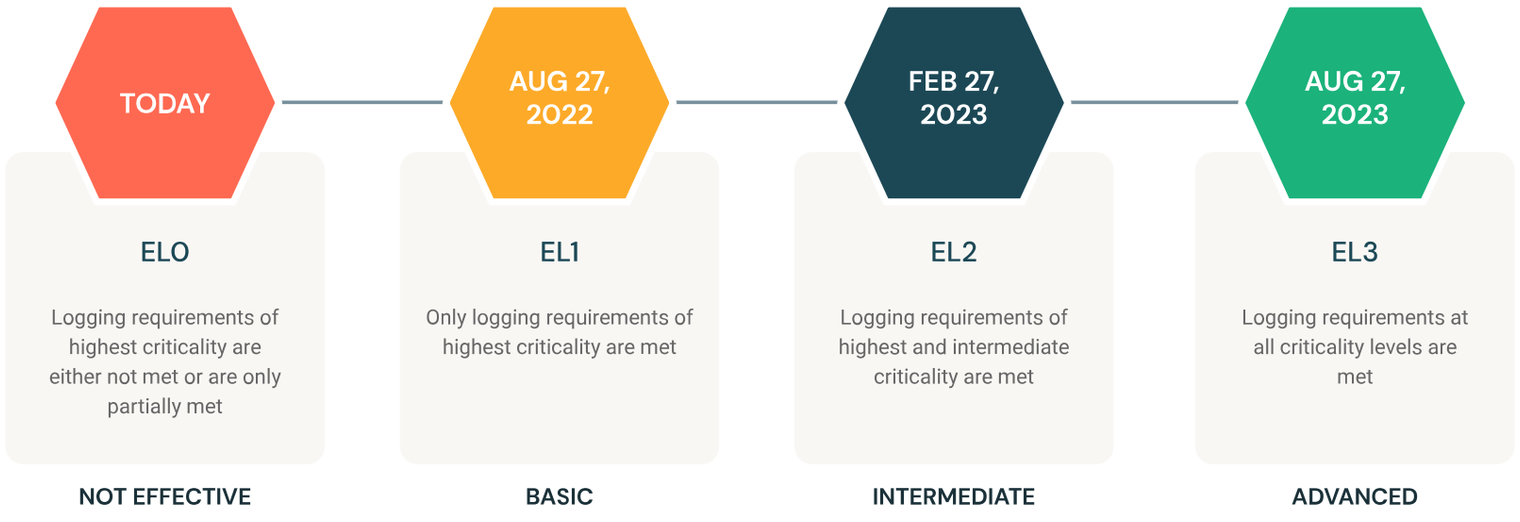
As the digitization of modern life continues full steam ahead, criminals and state-sponsored actors continue to target government systems. The impact of these attacks on the public sector range from the exposure of highly sensitive data to the disruption of services and the exploitation of back doors for future attacks — all resulting in financial costs. In 2021, the U.S. government faced costs of more than $4 million per cyber incident.

It comes as no surprise, then, that recent years have seen an amplified commitment to cybersecurity in the public sector. As government leaders look to new solutions, large portions of the federal budget are now devoted to leveraging data and AI to thwart cyberattacks. For example, government IT expenditure is expected to surpass $109 billion in 2022, up from $92 billion in 2021. In terms of budget allocation, the Department of Defense stands out as the primary recipient of federal cybersecurity spending, since the agency is responsible for protecting the United States from both offline and online attacks.

Furthermore, on August 29, 2021, the U.S. Office of Management and Budget (OMB) released a memo in accordance with the Biden administration's Executive Order (EO) 12028, Improving the Nation's Cybersecurity, which mandates that federal agencies adapt to today's cybersecurity threat landscape. The memo (M-21-31), adds complexity requiring federal agencies to meet each rising level of maturity using their existing cybersecurity budget. See the diagram below.

However, as is often the case, implementing new strategies and processes is easier said than done.

In this eBook, we'll take a closer look at the challenges associated with replacing the infrastructure of a legacy data analytics system, and how different branches of the public sector are solving them with Databricks.



The Biggest Challenge With Security Analytics

For many organizations on-premises security incident and event management (SIEM) technologies have been the go-to solution for threat detection, analysis and investigations. However, these legacy technologies were built for a world where big data was measured in gigabytes, not today's terabytes or petabytes. This means that not only are legacy SIEMs unable to scale to today's data volumes, they are also unable to serve the modern, distributed enterprise.

By now, the advantages of moving to the cloud are no secret to anyone. For public sector organizations, scalability, simplicity, efficiency and cost are absolutely essential components of success. Many public sector organizations are looking to cloud computing to make this possible — adding detection and response in the cloud to security team's responsibility.

Because legacy SIEMs predate the emergence of cloud, machine learning (ML) and AI in the mainstream, they're unable to address the complex ML- and AI-driven analytics needed for threat detection, threat hunting, in-stream threat intelligence enrichment, analytical automation and analyst collaboration.

In other words, legacy SIEMs are no longer suitable for the modern enterprise or the current threat landscape.

Counting the financial cost of legacy SIEMs

The financial cost of the continued use of legacy SIEMs continues to rise because most SIEM providers charge their customers based on the volume of data ingested. While some legacy technologies are available in the cloud, they're confined to a single cloud service provider. As a result, security teams have to employ multiple tools for detection, investigation, and response — or pay exorbitant egress charges for data transiting from one cloud provider to another. This causes operational slowdowns, errors driven by complexity, and inconsistent implementation of security policies.

A lack of support for multiple clouds also means an increase in maintenance overhead. Security staff members are often stressed because analysts have to learn different tools for different cloud platforms. For some, it also creates an implicit cloud vendor lock-in, meaning that security teams are unable to support missions because their tools are not portable across multiple cloud providers.

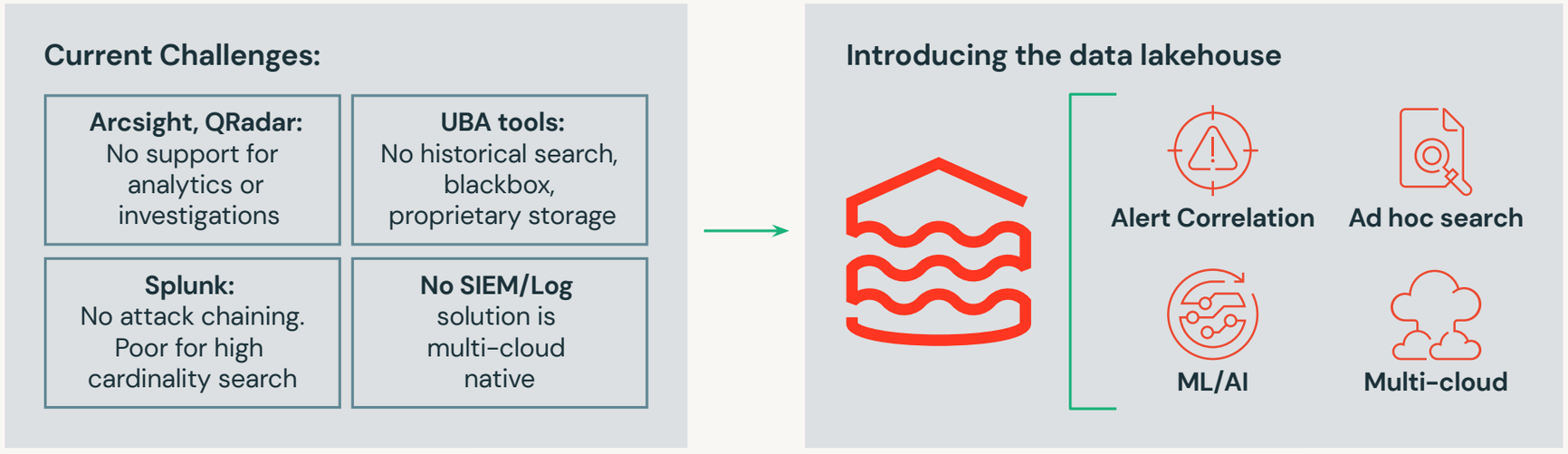
Collectively, these drawbacks to legacy SIEMs result in a much weaker security posture for the public sector.

databricks

Journey of SecOps: Destination Lakehouse

How did security analytics get to this point? In the early days there was a need to aggregate alerts from antiviruses and intrusion detection systems. SIEMs were born — built on data warehouses or relational databases. But as incident investigation needs emerged, those data warehouses weren't able to handle the volume and variety of data which led to the development of data lakes. Data lakes were cost-effective and scalable, but didn't have strong data governance and data hygiene — earning them the moniker of data swamps. Simply integrating the two tech stacks is really complicated because of varying governance models, data silos, and inconsistent use case support. Fast forward to today, where security teams now need AI/ML in a multi-cloud world.

Why choose one or the other? What if government agencies could have it all? The governance and transactional capabilities of the data warehouse, the scale and flexibility of a data lake, ML/AI from the ground up AND multi-cloud native deployments? This is a modern architecture called the lakehouse (data lake + data warehouse).



databricks

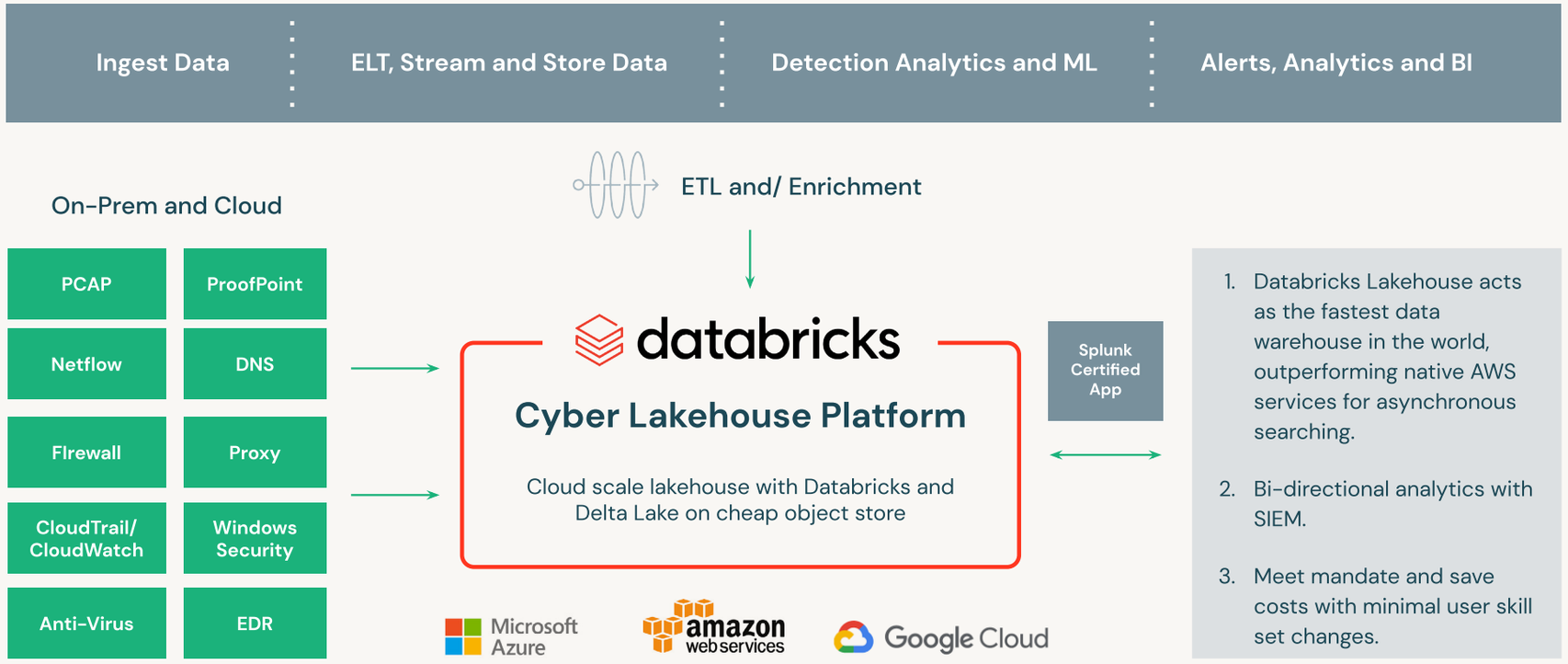
Why Databricks Lakehouse for Cybersecurity

Databricks introduced the first data lakehouse platform to the industry and today over 7,000 customers use it worldwide. With Lakehouse, government agencies ready to modernize their data infrastructure and analytics capabilities for better protection against cyber threats now have one cost-effective solution that addresses the needs of all their teams.

The Databricks Lakehouse Platform combines the best elements of data lakes and data warehouses, delivering the low-cost, flexible object stores offered by data lakes and the data management and performance typically found in data warehouses. This unified platform simplifies existing architecture by eliminating the data silos that traditionally separate analytics, data science and ML. It's built on open source, open data and open standards to maximize flexibility, and its native collaborative capabilities accelerate the ability to work across teams and innovate faster. Moreover, because it's multicloud-native, it works the same way no matter which cloud providers a public sector organization uses.

Databricks Cyber "Multi-Tier" Architecture

A lakehouse platform for scalable, real-time threat analytics



Lakehouse in the Public Sector

By unifying data with analytics and AI, Lakehouse allows federal agencies to easily access all their data for downstream advanced analytics capabilities to support complex security use cases. Lakehouse enables government security operations teams to detect advanced threats, accelerate investigations from days to minutes, and reduce human resource burnout through analytical automation and collaboration.

Along with a more modern architecture, the Lakehouse Platform includes Delta Lake, which unifies all security data in a transactional data lake to feed advanced analytics. The analytics and collaboration are done in notebooks, and security teams can use multiple languages — SQL, Python, R and Scala — in the same notebook. This makes it easy for security practitioners to explore data and develop advanced analytics and reporting using their favorite methods. Additionally, a separation of compute from storage means performance at scale without impacting overall storage costs.

Harnessing data for competitive military advantage

The U.S. Air Force implemented the Lakehouse to create the Visible, Accessible, Understandable, Linked and Trusted (VAULT) data platform. VAULT is the Air Force's network of cloud-based analytics tools designed to improve readiness and promote mission success. Prior to signing on with Databricks, the team behind the VAULT effort struggled to support their rapidly growing data sets with a legacy technology infrastructure. The legacy system was very difficult to maintain and almost impossible to upgrade, and did not facilitate collaboration across teams — ultimately creating data silos.

With the Lakehouse Platform, the U.S. Air Force has simplified their data architecture by eliminating old data silos, which in turn streamlined its analytics, data science and ML capabilities.

This was particularly game-changing once COVID-19 hit and employees had to work remotely. With the cloud-native Lakehouse, the Air Force was able to comply with social distancing requirements while achieving mission objectives. The openness of Lakehouse enabled the Air Force to push information to a classified environment. Additionally, because VAULT is hosted on Amazon Web Services (AWS), which is accredited at the highest security impact levels, the Air Force has a capability that grew tenfold in 12 months.

The Air Force is also leveraging Databricks' partner Immuta's data policy enforcement platform to automate granular access and privacy controls, with both attribute and purpose-based access control models natively integrated within the Lakehouse. This architecture has streamlined data sharing and data access, and enabled rapid modernization to empower users via self-service analytics driven by Databricks.

"When I look at where we've come in our department, we're really about driving data innovation and empowering our department to harness data for competitive military advantage," said Eileen Vidrine, chief data officer for the Department of the Air Force. "t's using and sharing data as a catalyst for innovation and driving capabilities, providing our airmen and guardians the capability, in a self-service environment, to leverage data to drive insights."

databricks

Advantages of a Lakehouse for Public Sector Agencies

A cost-efficient upgrade

Databricks customers only pay for the data they analyze, not for what they collect. This means that security teams can collect any amount of data without worrying about ingest-based pricing, and only pay for the data that's actually used for analysis — for example, an incident investigation or a data call for an audit. This pricing model enables security teams to collect data that was previously out of reach, such as netflow data, endpoint detection and response data, and application and services data.

Further, Databricks is a fully managed service, meaning that security teams don't have to pre-commit to hardware capital expenditures. With no hardware to manage and no big data implementations to maintain, security teams can significantly reduce their management and maintenance costs.

Cloud-agnostic

Databricks is cloud-native on AWS, Microsoft Azure and Google Cloud Platform (GCP). This creates freedom for the security teams to use whatever cloud provider they like. Additionally, security teams can acquire and maintain operational consistency across all providers when they have multiple cloud footprints. This enables consistent policy implementation, reduced complexity for staff and increased efficiency.

For public sector agencies that support other agencies, Databricks enables faster detection, investigation and response because agencies can reuse analytics across the three major clouds through a unified platform that centralizes data for easy sharing and fosters collaboration across teams.

Enterprise security and 360° administration

The Lakehouse Platform is easy to set up, manage, scale and, most importantly, secure. This is because Lakehouse easily integrates with existing security and management tools, enabling government users to extend their policies for peace of mind and greater control.

With multicloud management, admins and data teams get a consistent experience across all major cloud providers. This saves valuable time and the resources required to upskill talent on proprietary services for data, analytics and AI.

Admins are also able to give team members a range of security permissions that come with thorough audit trails. This allows admins to quickly spin up and wind down collaborative workspaces for any project and to manage the project end to end — from enabling user access and controlling spend to auditing usage and analyzing activity across every workspace to enforce user and data governance.

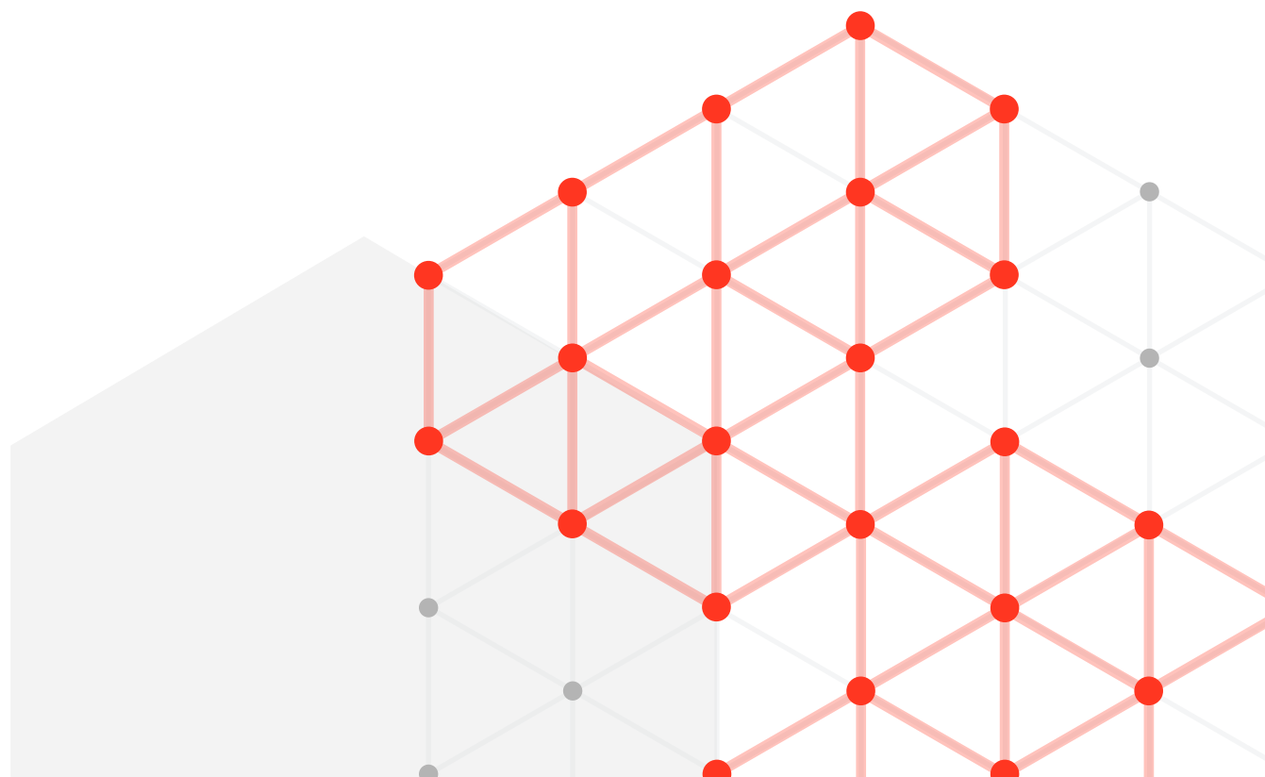
databricks

Lakehouse + SIEM: The Pattern for Cloud-Scale Security Operations

According to George Webster, head of cybersecurity sciences and analytics at HSBC, Lakehouse + SIEM is the pattern for security operations. What does it look like? It leverages the strengths of the two components: Lakehouse for multicloud native storage and analytics, SIEM for security operations workflows. For Databricks customers like HSBC, there are two general patterns for this integration that are both underpinned by what Webster calls, the cybersecurity data lake with Lakehouse.

In the first pattern, Lakehouse stores all the data for the maximum retention period. A subset of the data is then sent to the SIEM and stored for a fraction of the time. This pattern has the advantage of allowing analysts to query near-term data using the SIEM while having the ability to do historical analysis and more sophisticated analytics in Databricks. It also lets them manage any licensing or storage costs for the SIEM deployment.

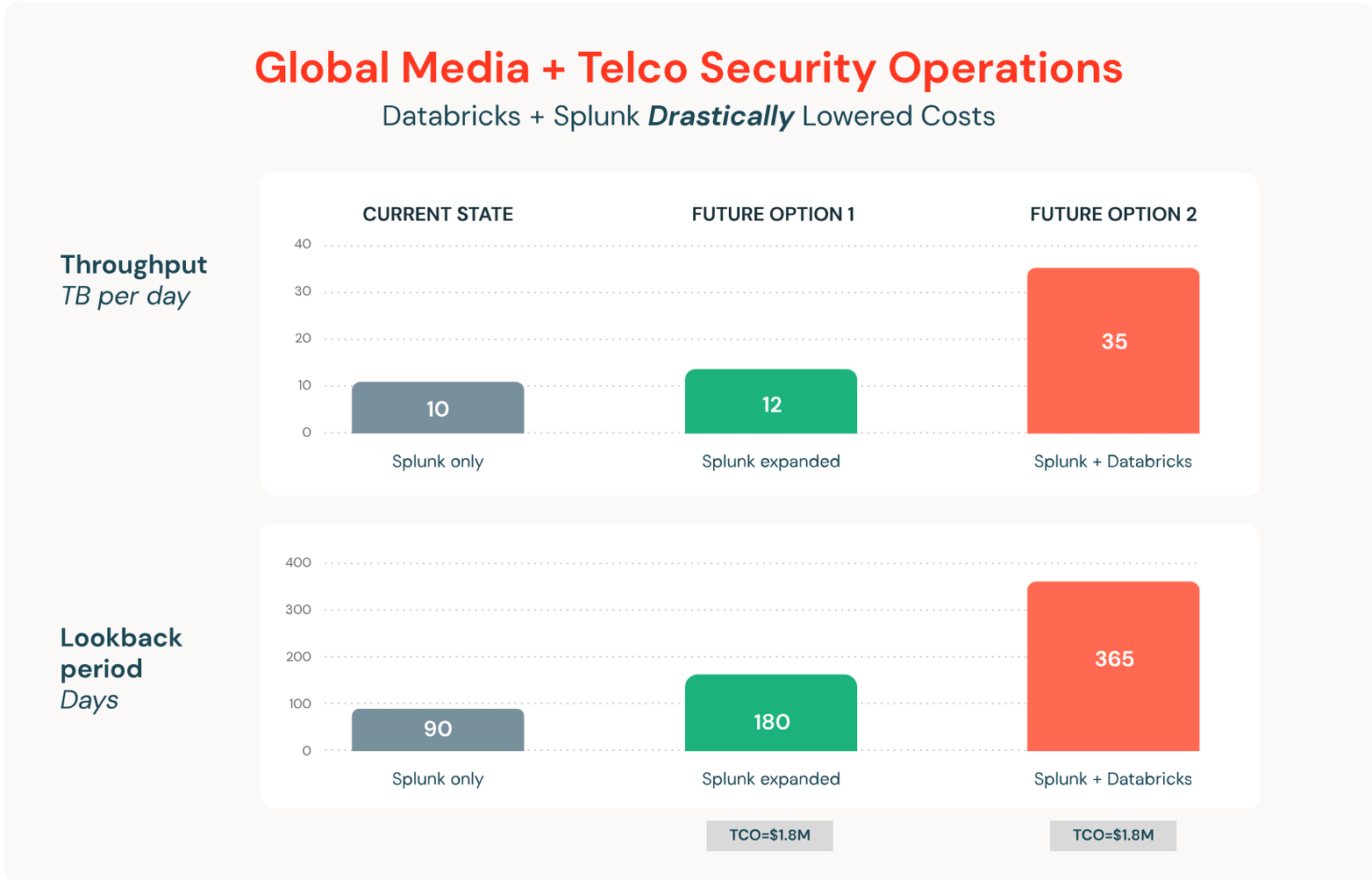
The second pattern is to send the highest-volume data sources to Databricks — for example, cloud-native logs, endpoint threat detection and response logs, DNS data and network events. Low-volume data sources such as alerts, email logs and vulnerability scan data — go to the SIEM. This pattern enables Tier 1 analysts to quickly handle high-priority alerts in the SIEM. Threat-hunt teams and investigators can leverage the advanced analytical capabilities of Databricks. This pattern has a cost benefit of offloading processing, ingestion and storage from the SIEM.



databricks

Databricks + Splunk: a case study in cost-savings

Databricks integrates with your preferred SIEM, like Splunk, and our Splunk-certified Databricks add-on can be used to meet both OMB M-21-31 and SOC needs without changing the user interface. While this example features a global media telco's security operation, the expanded lookback and data ingestion matches the cybersecurity needs of government agencies. The telco organization grew throughput from 10TB per day with only 90 days look back, to 35TB per day with 365 days lookback using the Databricks SIEM augmentation. Despite a 250% increase in data throughput and more than quadrupling the lookback period, the total cost of ownership, including infrastructure and license, remained the same, saving 10s of millions per year in cloud costs.



databricks

Common Use Cases

From the Federal Data Strategy to the National AI Initiative Act, it's clear that the U.S. federal government is focused on modernizing its data analytics and warehousing capabilities. The Databricks Lakehouse Platform brings a new level of empowerment to federal agencies, allowing them to unlock the full potential of their data to deliver on their mission objectives and better serve citizens. Common use cases include:

- **Threat hunting:** Empower security teams to proactively detect and discover advanced threats using months or years of data
- **Incident investigation:** Gain complete visibility across network, endpoint, cloud and application data to respond to incidents
- **Phishing threat detection:** Uncover social engineering attacks that are often used to steal user data, including log-in credentials and credit card numbers
- **Supply chain monitoring:** Leverage ML to identify suspicious behavior within your software supply chain
- **Ransomware detection:** Scope the impact and spread of ransomware attacks to inform complete mitigation and remediation
- **Credentials-abuse detection:** Identify and investigate anomalous credential usage across your infrastructure
- **Insider-threats detection:** Find and respond to malicious threats from people within an organization who have inside information about security practices, data and computer systems
- **Network traffic analysis:** Examine real-time network availability and activity to identify anomalies, vulnerabilities and malware
- **Analytics automation:** Automatically contextualize and enrich multiple streaming and batch analytics to accelerate analyst workflows and decision making

Getting Started With Databricks for Cybersecurity

Getting up and running on Databricks to address your cybersecurity needs is easy with our Solution Accelerators. Databricks Solution Accelerators are highly optimized, fully functional analytics solutions that provide customers with a fast start to solving their data problems.

- Cybersecurity analytics and AI at scale with Splunk and Databricks: Rapidly detect threats, investigate the impact and reduce risks with the Databricks add-on for Splunk
- Threat detection at scale with DNS analytics: Recognize cybercriminals using DNS, threat intelligence feeds and ML

Databricks Solution Accelerators are free. Join the hundreds of Databricks customers using Solution Accelerators to drive better outcomes in their businesses.

If you'd like to learn more about how we are helping the public sector securely leverage data and AI, please visit us at dbricks.co/federal or reach out to us at cybersecurity@databricks.com

databricks

About Databricks

Databricks is the data and AI company. More than 7,000 organizations worldwide — including Comcast, Condé Nast, Acosta and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark,™ Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

Get started with a free trial of Databricks and start building data applications today

START YOUR FREE TRIAL

To learn more, visit us at:

Federal Solutions

databricks