International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I6P104

Volume 71 Issue 6, 25-29, June 2023

Original Article

Enhancing Cybersecurity Through the Unification of Data Analytics, Artificial Intelligence, and Machine Learning in Big Data Cloud Environments: A Databricks Lakehouse Approach

Mayur Katariya[1], Digan Parikh[2]

[1]*Delivery Solutions Architect, Databricks, CA, USA.*
[2]*Solutions Architect, Databricks, CA, USA.*

Received: 28 April 2023 Revised: 25 May 2023 Accepted: 09 June 2023 Published: 30 June 2023

Abstract *- "Data without analytics is dormant, the cloud without security is vulnerable, and AI without data is blind. Together, they form an unstoppable force in the realm of cybersecurity." The cybersecurity industry faces numerous challenges in combating the ever-evolving threat landscape. This scholarly article aims to shed light on these challenges and explores how Databricks' lakehouse cloud architecture offers potential solutions. The article provides a comprehensive detail of the current state of the cybersecurity industry and maturity curve and highlights the pressing issues that organizations encounter. The article focuses on the utilization of Databricks' lakehouse architecture as a means to address these challenges. Furthermore, the paper delves into the key features and capabilities of Databricks' lakehouse architecture that enable organizations to efficiently analyze and govern vast amounts of security data, detect anomalies, and identify potential threats in real time. The paper concludes by drawing on real-world examples and industry case studies. This paper offers a detailed overview of the real-world cybersecurity architecture examples, data flow and how different companies are embracing the lakehouse architecture to power their use cases. By unifying the combined power of data analytics, artificial intelligence (AI), and machine learning (ML), the cloud lakehouse architecture helps organizations to break silos, derive actionable insights, offer a promising approach for bolstering cybersecurity defenses, and enhance their proactive cybersecurity strategies at scale on cloud.*

Keywords *- Databricks lakehouse, Cloud, Bigdata, Analytics, Artificial Intelligence(AI), Machine Learning(ML), Cybersecurity.*

1. Introduction

Cybersecurity has become an increasingly important concern in the digital age. As technology continues to advance, cyber threats are becoming more sophisticated, and the challenge of securing data has become more complex, as shown in the list of attacks.[1] Data breaches have affected companies and organizations of all shapes, sizes, and sectors, and they are costing US businesses millions in damages[2]. The technology maturity curve is constantly evolving, and businesses need to stay up to date with the latest cybersecurity solutions to protect their assets.

When we think about the cybersecurity maturity curve (see Fig 1), most enterprises are in the early stage of this curve. Some enterprises have figured out how to merge reactive insights with predictive insights; however, most enterprises still struggle to move from hindsight (reactive insights) to foresight (Predictive insights and proactive actions). They are so focused on alerts, reports, batch detection, and ad-hoc queries that they often seem to forget about the preventive AI & ML-based approach that will give them a competitive edge over others.

2. Challenges in Cybersecurity

There are four common challenges in tackling cybersecurity operations in the cybersecurity industry. These challenges include:

2.1. Costs

Security data is coming from many different sources with an urgency to retain everything in different locations. This leads to an explosion in storage costs. Not only that, but some cyber teams also have a vendor sprawl. They spend a lot of time on administration, governance, licensing, management of ETL pipelines, data, and process overhead.

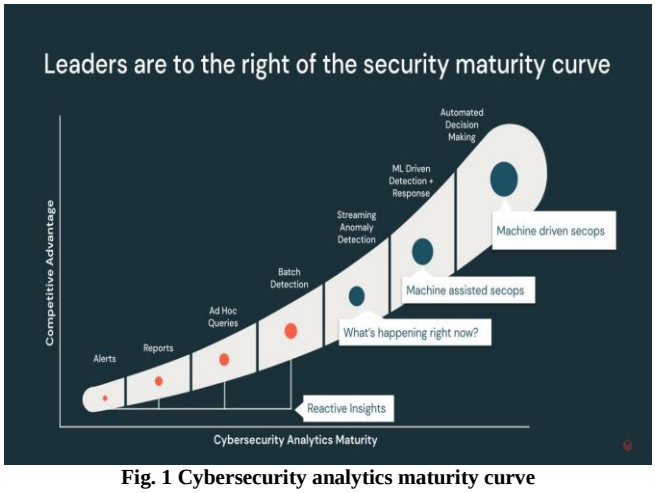

Fig. 1 Cybersecurity analytics maturity curve

This is a big pain point, especially since enterprises want to lower their total cost of ownership (TCO) and want to become more profitable.

2.2. Scale

Many large enterprise security teams struggle to process large data columns from heavy data sources. There are limited investigation capabilities and high hardware costs with non-elastic licensing from vendors. Often, traditional SIEMs struggle to track and process high volume and high cardinality data for threats. This leads to some threat hunting and security detection challenges like stateless hunting, non-performant search queries or limited data scans.

2.3. Artificial Intelligence & Machine Learning

With bigdata, incorporating AI & ML into security projects is a must to gain a competitive edge. Despite their intentions and efforts, picking the right machine-learning tools and technologies is a must. The AI & ML industry is highly fragmented, with vendors specializing in different parts of the ML lifecycle, making it difficult to find a solution that works. For example, you see modern Data Warehouse solutions try to fit into the ML lifecycle but fail in execution. These vendors end up creating incompatible, limiting, and proprietary solutions and lack full-scale support with common ML Ops methodologies such as AutoML, Feature store, Model Registry, Model Monitoring, GPU support and so on.

2.4. Centralization with Governance

Data is split across many different data stores, creating the challenge of consolidating data into one location. Within an enterprise, you might have data stored in traditional SIEMs, Cloud-based SIEMs and data lakes, making it almost impossible to create a single source of truth. This not only ends up creating data silos but also makes it super difficult to integrate and govern the data with ease.

3. How Databricks Lakehouse Solves these Challenges and Empowers Cybersecurity Customers?

3.1. First, let us try to Understand what is Lakehouse Architecture

A Cybersecurity Databricks Lakehouse (See Fig 2) is a modern data architecture that combines the features of a data lake and a data warehouse. It enables organizations to store, manage, orchestrate, and analyze large volumes of cybersecurity data in a centralized, scalable, governed & performant manner, making it easier to identify and respond to security threats in real time.

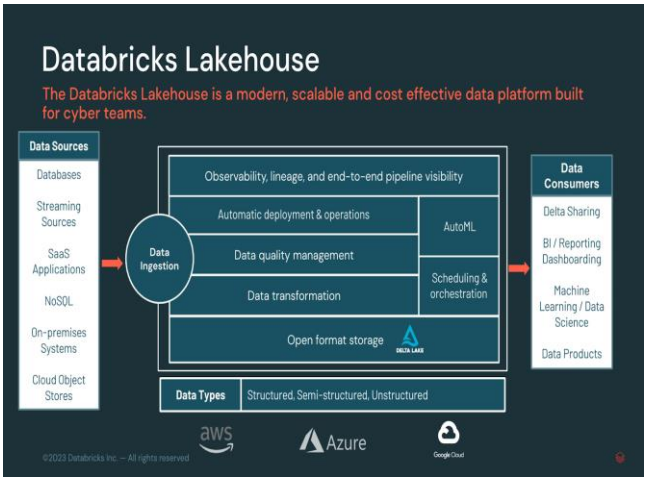

Fig. 2 Databricks lakehouse architecture

Lakehouse is user agnostic - no matter what persona (Threat researcher, SOC analyst, data scientist or ML engineer), you can use the data lakehouse with the same paradigm to achieve the goal.

The Lakehouse platform simplifies the complexity of building and maintaining pipelines and running all types of workloads in one single platform. All types of personas, whether it is a SOC analyst, threat researcher, data engineer, data scientist or ML engineer, can access the same platform using the choice of their interface. Security analysts (SOC) & threat researchers can search for the data natively using a SQL interface using various dashboards or can be integrated with their choice of BI tools. Data Engineers can build data pipelines using workflows. Data scientists & ML engineers can use collaborative notebooks for interactive and advanced analysis using the native Python, R, Java, and Scala support.

3.2. Now, let us Relate how these Lakehouse Capabilities Address the Challenges Highlighted Earlier

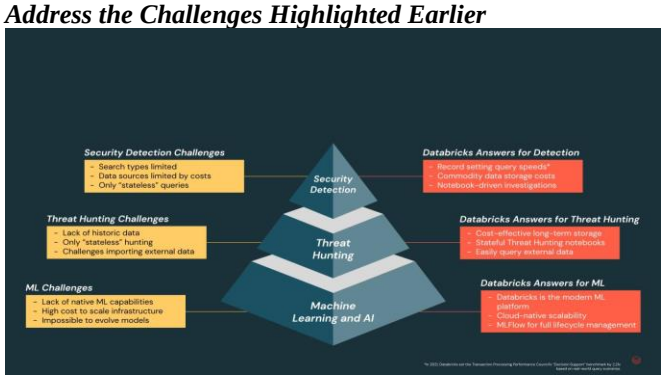

Fig. 3 Cybersecurity lakehouse addresses all challenges

3.2.1. Costs

With Databricks Lakehouse, the enterprise has custody of the data. This opens up the enterprise to make flexible data decisions in the future and avoid vendor lock-in. Data is stored centrally, in cheap object storage in one or multiple cloud locations - S3, ADLS and GCS. This means that data can be ingested from multiple types of data sources in any format - structured, semi-structured or unstructured. With this, Databricks provides a separation of computing and storage. Cybersecurity data can sit in cheap cloud storage without any computing running. Databricks uses Apache Spark, an open-source unified distributed analytics engine for large-scale data processing for computing. This further helps avoid vendor lock-in, as you can migrate your workloads to any system that supports Spark. With Databricks Serverless capabilities, the management, scaling, and optimization of computing happen instantaneously within a matter of seconds, and organizations do not need to worry about their valuable compute resources running idle, thereby helping lower TCO.[3]

Enterprises nowadays have a huge emphasis on automatic deployment and operations to reduce costs. Cybersecurity teams can orchestrate data pipeline workflows with a scheduler, providing end-to-end orchestration capabilities within Databricks. To add to the automation, enterprises can leverage terraform to treat infrastructure as code. This ensures templates are created once and reused multiple times per demand, resulting in significantly lower TCO.

3.2.2. Scalability

In terms of scale, Databricks Lakehouse provides an elastic compute scale as needed. This allows threat hunters to perform searches over petabytes of historical data with built-in shared GIT repositories for threat hunting and investigation notebooks. This scalability makes it easier to manage and analyze data in real time as data grows without the need for frequent data migrations. Moreover, Delta Lake [4] supports ACID transactions, which ensure data consistency and integrity. This means that transactions are Atomic (all or nothing), Consistent (data meets pre-defined rules), Isolated (transactions do not affect each other), and Durable (committed data is permanent). This feature is essential for ensuring the integrity of security-related data, such as logs, audit trails or GDPR & CCPA [5]compliance reasons. Check out the security and trust center[6]

Fig. 4 Lakehouse solves detection challenges at scale

3.2.3. Artificial Intelligence & ML for real-time and Predictive Analytics

A cybersecurity Lakehouse can leverage advanced analytics tools and techniques like machine learning and artificial intelligence to identify potential threats, detect anomalies, and improve overall security posture. Data Scientists and ML engineers can use the Databricks DS & ML workbench for detections, co-pilot investigations & remediations and risk modeling. It offers full, end-to-end MLOps[7] features. AutoML allows users to quickly generate baseline models and notebooks with the auto-generated code as an output. ML experts can accelerate their workflow by fast-forwarding through the usual trial-and-error and focusing on customizations using their domain knowledge, and citizen data scientists can quickly achieve usable results with a low-code approach. Machine learning can be initiated automatically, and its features and models can be saved for reuse and production migration. MLflow introduces experiment tracking, projects, model portability and model lifecycle management with governance.

Moreover, by providing real-time access to cybersecurity data, Databricks Lakehouse enables security teams to quickly identify and respond to threats using these AI/ML capabilities as they occur, reducing the impact of security incidents.

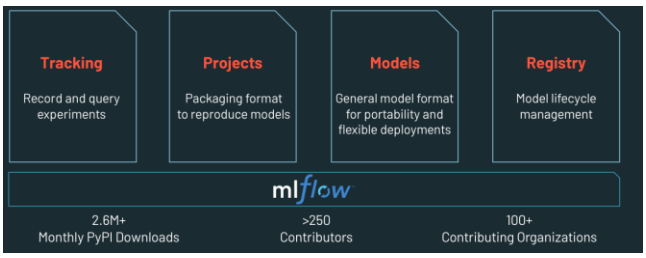

Fig. 5 MLFlow

3.2.4. Centralization with Governance

A cybersecurity lakehouse provides a centralized location to store all cybersecurity data, including network logs, security events, and user behavior data. By keeping all cybersecurity data in one place, it makes it easier to access data, perform analytics, generate reports, and maintain data quality, data traceability and governance. Data can be transformed upon ingestion with Spark SQL, PySpark, R, and Scala APIs, while data quality is maintained with schema enforcement and data validations.

A commonly requested feature from cybersecurity teams is data versioning. With Delta Lake, every change made to the data is recorded and tracked and can revert in time to visit older versions of data. This feature is important for cybersecurity solutions as it enables auditing and traceability of data changes, making it easier to identify and investigate past potential security breaches that are already in the logs or alerts. In terms of governance, Unity Catalog[8] provides centralized fine-grained access control, auditing, lineage, and data discovery capabilities which are critical for centralized data governance. Data lineage helps cybersecurity organizations be compliant and audit-ready, thereby alleviating the operational overhead of manually creating the trails of data flows for audit reporting purposes. To summarize, it can have data flow from multiple sources into the data lake in a governed and audited way (Fig 5).

How are companies using Databricks Lakehouse to power their cybersecurity use cases on the cloud?

Below Fig 6 is a high-level overview of the Lakehouse as it applies to Cyber Analytics and how different cybersecurity companies leverage Databricks Lakehouse for their use cases. The Lakehouse platform provides one unified platform that automates the complexity of the building, maintaining, and governing the data pipelines and running ETL workloads and machine learning directly on a data lake so that data professionals can focus on quality and reliability to drive valuable insights and analytics. Below Fig 6 showcases a sample process that can help you drive valuable insights and analytics.

- Data Ingestion - Wide variety of data coming from sources is ingested into the delta lake using cloud storage through event streaming services like kafka, kinesis and event hubs.
- Medallion Architecture [9] - A multi-hop architecture is used to load the data into bronze, silver, and gold layers.
    - ✓ The bronze layer contains raw datasets, which include the “as-is” data from the source systems.
    - ✓ This bronze data is enriched with additional metadata, and selective fields are extracted and schematized in the silver tables. These silver tables can be used for various threat model training and detection.
    - ✓ Ultimately, by joining various silver tables, different gold tables are created. This includes aggregated data and metrics to help with various business insights depending on the use cases. In this case, alert data can be aggregated by various event types and generate reports that can be used for SIEM Alerts, incident management and investigations.

- Unity catalog is used for fine-grained access control and lineage that can help with easy source-to-target tracking and auditability to help with faster impact analysis.

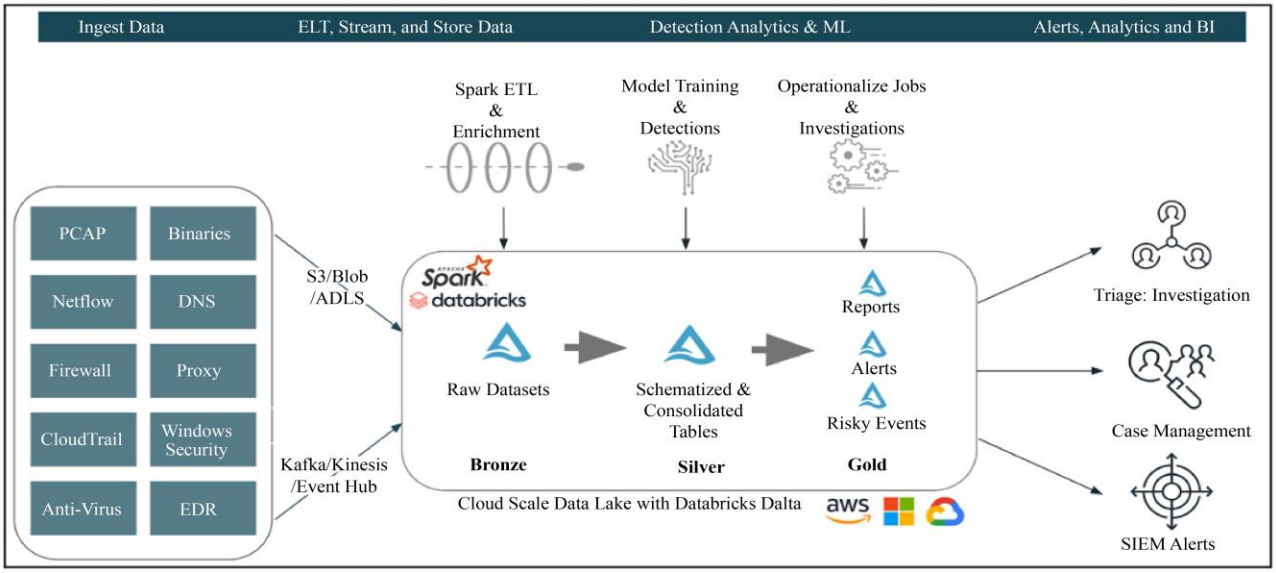

Fig. 6 Databricks lakehouse architecture for cyber analytics

Below are some examples and case studies on how different companies leverage Databricks Lakehouse for their security-related use cases.

Apple uses the Databricks platform for threat detection and response at scale. In fact, Databricks Delta technology was born because of a collaboration between Databricks and Apple due to this threat detection use case.[10]

HSBC uses the Databricks Lakehouse platform to detect and respond to threats across 115 TB/day by offloading expensive legacy SIEM. As a result, they were able to save more than $10M/Year in savings.[11]

Akamai[12], one of the leading content delivery network providers, is able to stream massive amounts of data and meet the strict SLAs it provides with its real-time analytics services to customers by leveraging Delta Lake and the Databricks Lakehouse Platform for the web analytics tool.[13]

Abnormal security reduced email-based threats through behavioral AI/ML & Analytical automation using Databricks Lakehouse. This resulted in 30% of productivity gains, $10M in revenues and cost savings of more than 40% reduction in TCO.[14]

4. Conclusion

As proven by these companies, Databricks Lakehouse provides the organization with a centralized, unified platform and scalable data architecture on the cloud leveraging powerful analytics, machine learning and AI capabilities that can help the enterprises to break the silos, improve their cybersecurity posture, reduce costs, and gain valuable insights into their security operations.

Acknowledgments

Lipyeow Lim, Cybersecurity SME and Technical Director, Cybersecurity at Databricks, for reviewing the article.

References

[1] Aaron Drapkin, Data Breaches That Have Happened in 2022 and 2023 so far, 2023. [Online]. Available: https://tech.co/news/data-breaches-updated-list
[2] Jade Artry, Average Data Breach Cost for US Businesses Almost $10 Million, 2023. [Online]. Available: https://tech.co/news/average-us-company-data-breach-cost-10-million
[3] Databricks Website. [Online]. Available: https://www.databricks.com/blog/announcing-general-availability-databricks-sql-serverless
[4] Databricks Website, what is Delta Lake?, 2023. [Online]. Available: https://docs.databricks.com/delta/index.html
[5] Databricks Website, GDPR & CCPA. [Online]. Available: https://docs.databricks.com/security/privacy/gdpr-delta.html
[6] Databricks Website, Compliance Security and Trust Center. [Online]. Available: https://www.databricks.com/trust#compliance
[7] Databricks Website, MLOps. [Online]. Available: https://www.databricks.com/resources/ebook/the-big-book-of-mlops
[8] Databricks Website, Unity Catalog. [Online]. Available: https://docs.databricks.com/data-governance/unity-catalog/index.html
[9] Databricks Website, Medallion-Architecture. [Online]. Available: https://www.databricks.com/glossary/medallion-architecture
[10] Spark + AI Summit, 2018 Vimeo Website. [Online]. Available: https://vimeo.com/274267634
[11] Data and AI Summit, Accidentally Building a Petabyte-Scale Cybersecurity Data Mesh in Azure With Delta Lake at HSBC, YouTube Website, 2022. [Online]. Available: https://www.youtube.com/watch?v=G9x-1s-1TJI
[12] Data and AI Summit, Akamai Interactive Analytics on a Massive Scale Using Delta Lake, YouTube Website, 2022. [Online]. Available: https://www.youtube.com/watch?v=_K9gIOsP4Os
[13] Databricks Website Customers – Akamai. [Online]. Available: https://www.databricks.com/customers/akamai
[14] Databricks Website Customers - Abnormal Security. [Online]. Available: https://www.databricks.com/customers/abnormal