Data quality:
The security differentiator

HOW GOOD, CLEAN SECURITY DATA ENABLES YOU TO HARNESS MORE CONNECTED INSIGHTS

DataBee®
A COMCAST COMPANY

databee.ai

Table of contents

The security **data landscape**

The lines between “data” and “security data” are being erased. We’re well into the new roaring 20s, and businesses and organizations are quickly learning that their ability to capitalize on their data — to turn raw and sometimes unmanageable data into actionable and connected business intelligence (BI) — is possibly the single most important weather vane when faced with either the head winds of opportunity or risks to the business.

While some business leaders and users scream that there is not enough data, the fact is, it’s just the opposite. There’s so much formless data, the word has almost lost its meaning. Data today is as prolific as dirt. Enterprises and large organizations ingest and digest mountains of the stuff every day. Conceptually, we know that “there’s gold in that data,” but only when it’s in a useable form can it be mined to expose security gaps, identify threats or vulnerabilities, or provide deeper insights, control, and visibility.

By unlocking the power of data, enterprises can...

Examine the **past.**

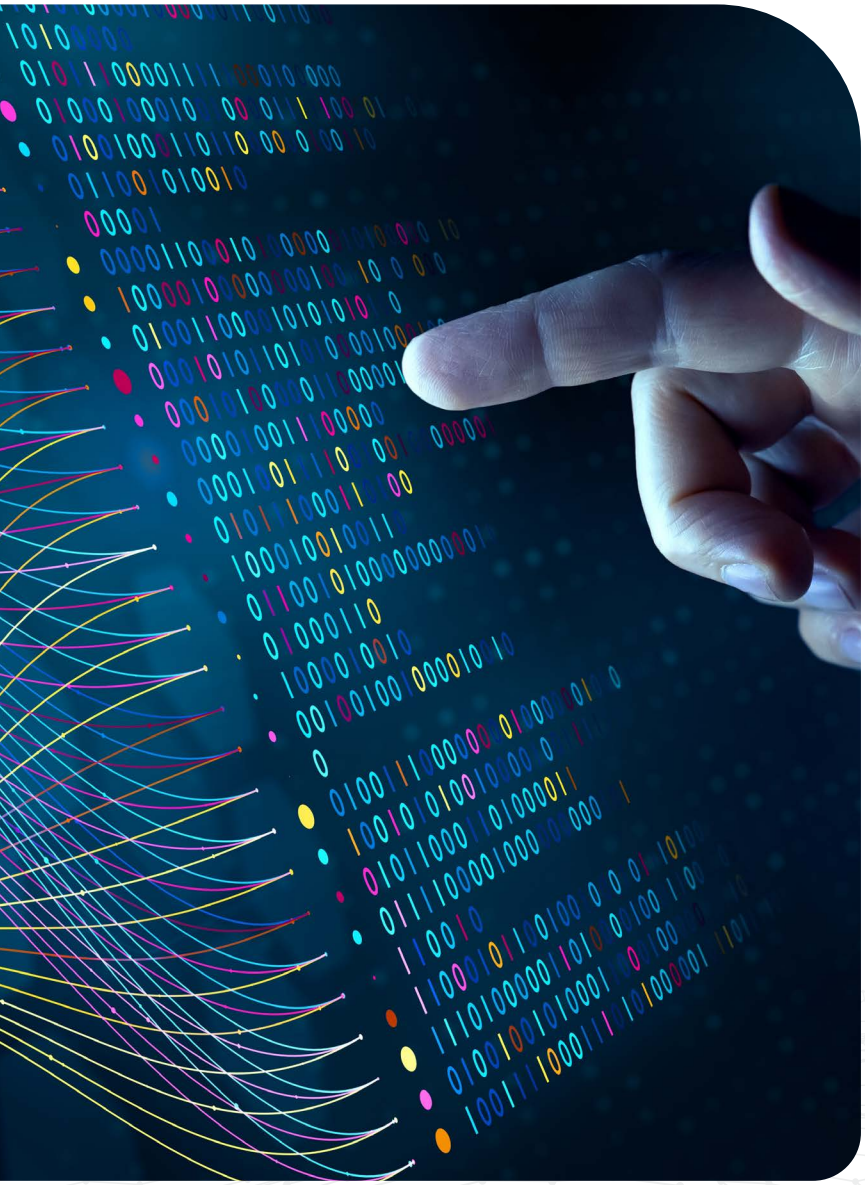
React to the **present.**

Chart the course to the **future.**

A diagnosis of “data dysmorphia”

It’s hard to know if you’re in great shape or not, data-wise. The question for organizations might start with “what do we know?” but if deep insights are that elusive, the question expands quickly into “how do we know how much we don’t know?”

Data analysis, and the creation of insights and narratives around what data is telling us, takes a lot of work. Given common challenges like siloed security controls, the limitations of next-gen security information and event management (SIEM) platforms, and incomplete configuration management databases (CMDBs) — all producing or working from fragmented, inconsistent, and raw data — it’s work that’s rife with unnecessary and duplicative digging. **The path forward starts with solving for data quality and data completeness.** Organizational challenges aside, by capturing data at or near the source, normalizing and enriching it to create a common wellspring for every business unit to work from, a company can accelerate time-to-value and turn the burden of big data into benefits that can help it truly differentiate.



How data quality **pays off**

Better data can help you gain confidence in your ability to protect your organization

The term GIGO — garbage in, garbage out — has been with us since the early days of computing, and of course, it does not just apply to code but to data as well. The data dilemma is just more pronounced when confronted with the sheer volume of today's data-centric ecosystem. And unfortunately, this is no different in the world of security data. On top of that, all the recent hype around AI, especially generative AI, has brought renewed attention to GIGO. Be wary of the organizational leader that invests in AI without first focusing on the data that powers it.

When you can establish high-quality data early in the data pipeline, normalizing the data format, centralizing the data collection and enrichment, cleansing for accuracy, and applying approaches like entity resolution (a patent-pending technology from Comcast), you can unlock the true potential of security and business information. It can aid organizations in their efforts to:

- **Accelerate** time-to-value for security-related data tools.
- **Support** well-informed security decisions and incident reporting, modeling, and analytics.
- **Facilitate** compliance and regulatory reporting and communication between teams.
- **Foster** trust in the insights from tools and systems.

All of these benefits can be compounded when realized together: better security, operational and business outcomes, improved transparency — leading to greater confidence wherever you want to drive data-driven decisions. Maybe even more important is the potential to mitigate avoidable problems such as:

Adverse or delayed outcomes from stale or fragmented data

False positives that waste resources on wild-goose chasing or low-priority incidents

The inability to answer security, operational, or business questions due to the lack of data

Wrong decisions that might impact brand, reputation, financials, security, and more

Cost is always going to be a consideration on both sides of the equation: How can a business weigh the cost of improving data quality against the cost of a more laissez faire data discipline? In some instances, security and governance, risk, and compliance (GRC) teams can end up spending more time trying to confirm data quality than in actually leveraging said data for reporting. There are many potential benefits to ensuring that datasets are built with high-quality data, including:

Teams can be dedicated to solving real business issues using data instead of fixing bad data decisions.

Insights require fewer cycles of checks, rechecks, and other redundancies.

Storage costs are aligned with actual needs and not just as a “data landfill.”

Data scientists, GRC experts, and security analysts can spend more time as storytellers and less time as data manipulators.

The data ecosystem

Not all data is insightful, but all data has value

The **data pipeline** is not just one stream but represents all the data channels used by an organization, both upstream and downstream — for data producers and consumers.

1. Security data comes directly from security and IT tools and systems (e.g., endpoint detection and response (EDR), firewalls, intrusion detection and prevention systems (IDPS), Windows event logs, DNS traffic, etc.)
2. It can also be combined — and enriched — with other enterprise systems (like Workday or ServiceNow, for example) to add context.
3. Enriched data usually undergoes a transformation process in order to convert/cleanse/structure it for use.
4. Policy context, and organizational information, can give the needed insight to determine if a finding is actionable, and how to take action upon it.

A data fabric, when combined with a data lake or other cloud or on-premises storage options, offers an optimal way to facilitate this process flow. Together they help answer the questions “what is my data telling me, and how do I weave it into my operations?”

A security data fabric is the technological cloth from which your data can be integrated and sewn together to help you produce readily available BI. Data fabrics are a complementary and modular architecture designed to enhance existing capabilities of your tools while also promoting integrity and cost control by streamlining data management earlier in the pipeline. They improve the way tools and data are integrated and how data is normalized, enriched, governed, and ultimately accessed by authorized users who must place their trust in what they’re seeing.

- Data fabrics can reduce the amount of work done to move or copy data by facilitating an abstraction layer for faster data integration and enrichment.
- From a vendor standpoint, the data fabric “ties the room together” by supporting different computing frameworks and programming languages so that all users, from data scientists to application developers and security analysts, can get the useable data they need.

Security data fabrics can pay specific dividends across a company when it comes to security, risk, and compliance analytics efforts by elevating transparency and traceability, and by allowing more focus on security data and telemetry. The semantics of security tools can be difficult to decipher; security data fabrics function as a sort of universal translator that can not only streamline workflows but also help alleviate the "black box" challenges that can occur within more proprietary security transformations.

- Security data is extremely nuanced, and most organizations require specialized skill sets to understand how to use the logs coming from these systems.
- Security data fabrics enable the enterprise to identify and respond quickly, provide a comprehensive view of their security posture across controls, manage network operations, and more.
- Auditors often appreciate open architectures and solutions with lineage and traceability as opposed to traditional black box approaches to risk and compliance scoring.

Data fabrics can be used as accelerators for artificial intelligence and machine learning (AI/ML) tools becoming a critical part of the operational landscape. The GIGO mantra applies in particular to machine learning tools, and the data fabric helps to ensure that they are learning and growing based on trusted curriculum. When built on top of a cohesive data fabric capable of producing rich metadata, AI/ML applications can become trusted partners that increase in value to the entire enterprise.

A security data lake, or other cloud or on-prem repository, are destinations for your downstream data to flow into, and that your organization can draw from. As we've established, data can come in many forms and from myriad sources. A large organization may have more than one data lake, but the first step in taking control of your data-driven destiny is to centralize and provide access to data as much as possible (with governance policies set by your organization), where it can better serve as raw material for security and operational needs.

The data lifecycle

On its journey through the pipeline, data goes through a myriad of processes

Ingest: Import disparate data files from multiple sources into a common location for processing.

Extract: Collect raw data from ingested files for transformation.

Parse: Analyze and separate data into smaller components for easier processing.

Normalize: Cleanse and standardize data format and structure.

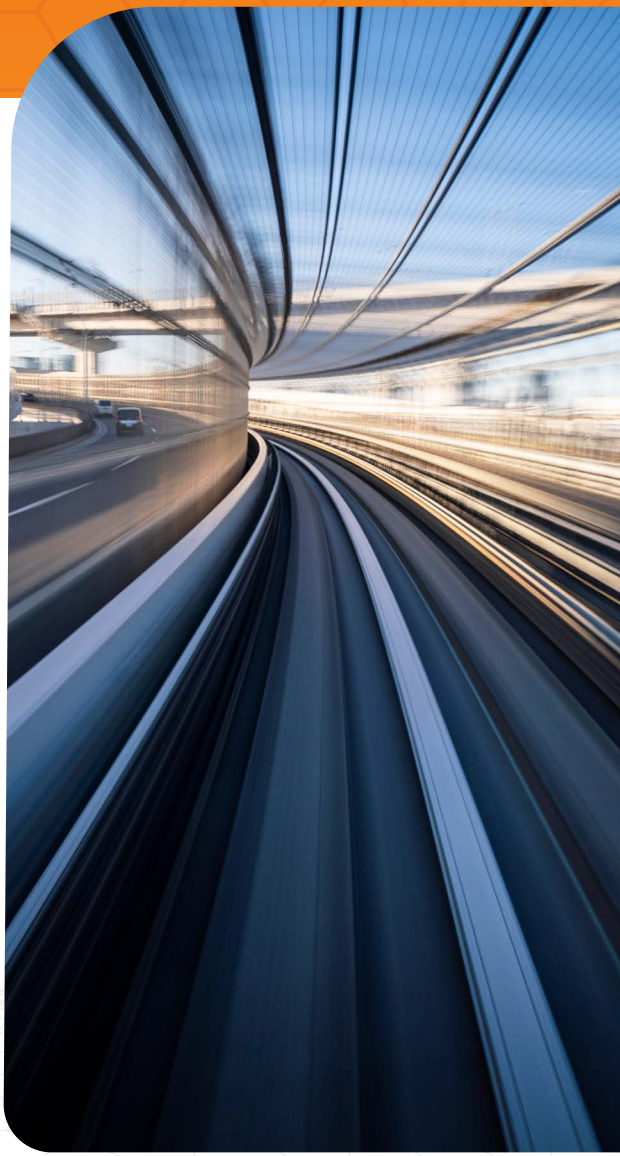
Flatten: Move semistructured data into a tabular format.

Enrich: Enhance existing data with supplemental data and entity resolution to improve reliability and value.

Store: Save the data within a secure infrastructure.

Governance: Apply policies pertaining to data access, sharing, and storage.

Provenance: Provide the lineage, including the data's origin.



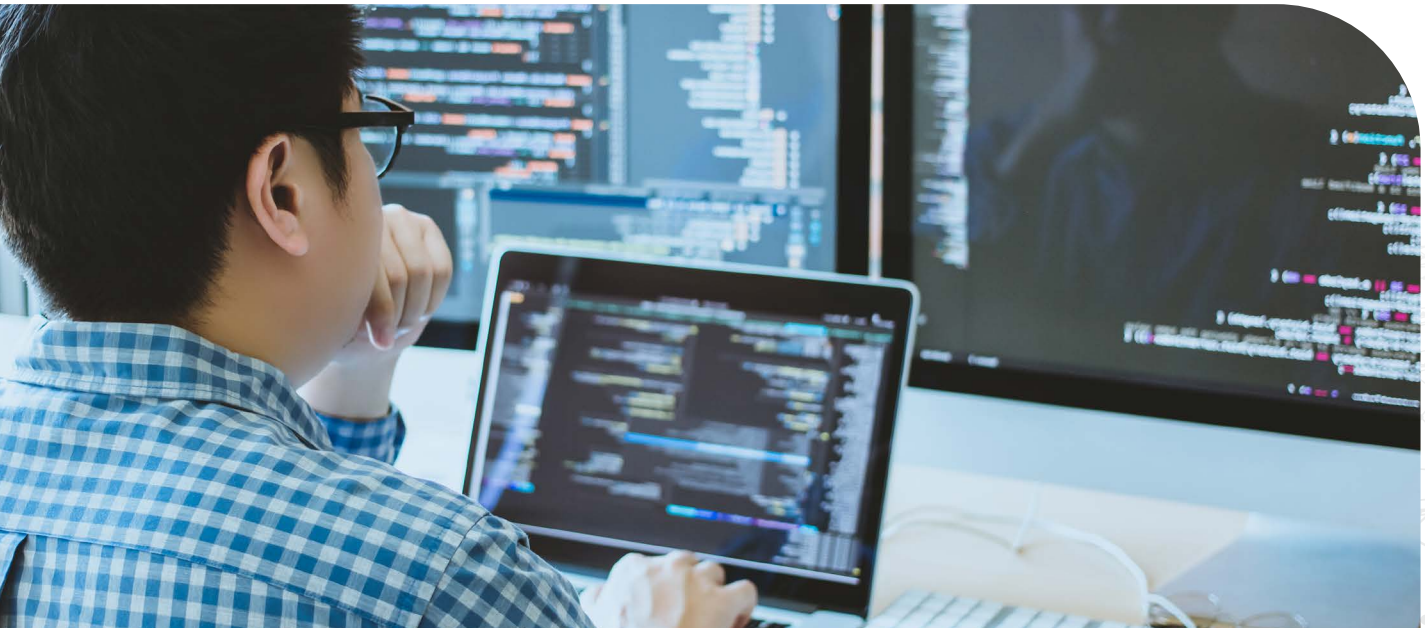
The elements of **good data quality**

Creating and executing your data quality game plan

It's clear that with disparate data sources and unique formats and syntaxes, elevating the quality and useability of data is not something that just happens organically. It takes a lot of work, much of it traditionally subject to manual processes. But a comprehensive, enterprise-wide plan can help automate and accelerate this transformation and lay the groundwork for continuous innovation of an organization's data maturity. Some primary considerations as you plan your future:

- Consider what "data of interest" is, define your scope, and create use cases that help visualize an optimal end state. Put simply, think of what you need your data to do, how your people will use it, and what outcomes you are driving toward.

- Confirm that identified data fields and sources are in alignment with not just your current use cases but also for near-future potential applications as well.



Careful, informed preparation is vital to implementing a new data architecture with minimal disruption to day-to-day operations. Your game plan should start with a clear understanding of the organization's existing components and how they interact.

1. Locate, capture, transform, and prepare, then store data of interest.
2. Determine what additional data would refine/ enhance analysis.
3. Decide where in the pipeline data will be parsed, normalized, and enriched.
4. Understand where data is shared today and how it can be shared optimally across the unique needs of your organization.
5. Have the AI/ML conversation. If downstream AI/ML implementation is part of current or future plans, now is the time to factor it in. Growing an effective, trusted, unbiased AI/ML discipline is a whole topic on its own and needs careful consideration.

All this talk about improving the useability and value of data begs the question **"what does this mean for my enterprise?"** Every organization has unique challenges and requirements that call for deep discovery and documentation. Some key questions include:

1. **Is the overall structure of downstream data aligned with expectations?**
   - What field structures (i.e., data formats) are compatible with BI tools or other downstream processes?
   - Are there proper and unique identifiers to facilitate data combination and insights at multiple levels of detail?
   - Is there a time-stamping field to indicate when data was ingested and processed?

2. **Is the downstream data truly accurate?**
   - Do field names accurately represent their corresponding value?
   - Are metrics calculated correctly and transparently?

3. **Is the downstream data truly complete?**
   - What controls and testing mechanisms are in place to ensure that relevant data isn't omitted during transformation? What needs to be added or augmented?
   - Is there deduplicated data that should not be deduped? What remedy should be in place?
   - Are incomplete datasets being joined, thus causing important data to be dropped?

"Shifting left" — Cybersecurity as a data priority

An enterprise's data maturity and its security acumen are inexorably tied, which means simply that one cannot evolve without the other. The need for more sophisticated threat detection and response, as well as better controls compliance monitoring, is transforming the relationship between organizations and their raw data. After all, in an information economy, data is no longer a differentiator; it's what you can do with it that matters, and that starts with elevating the quality and depth of the data relied upon by your teams and their tools. This new mindset is strengthening collaboration between security and IT groups as cybersecurity "shifts left" into a higher level of priority.

How DataBee® helps enterprises **get the most out of their data**

What can your data be?

Comcast NBCUniversal is a Fortune 33 company that produces and manages terabytes of data from diverse sources daily.

Comcast actively safeguards this ever-growing data landscape by leveraging a security, risk, and compliance data fabric that merges disparate data sources and feeds with organizational and business data. The result is proactive, enhanced threat visibility, detailed traceability, and faster time-to-value.

DataBee, a Comcast Company, inspired by the data fabric used within Comcast, focuses on the entire security data pipeline — giving you data in a clean, usable format before applying analytics. DataBee is ready for use by threat hunters, data engineers and scientists, SOC analysts, compliance and audit specialists, and incident responders.

VISIT US TO LEARN MORE: DATABEE.AI →

DataBee benefits:

- Use data to drive cross-departmental collaboration.
- Accept different data formats, then transform them to reside in one location.
- Make the dataset usable by disparate security teams answering different questions as a result of DataBee's use of the Open Cybersecurity Schema Framework (OCSF) and the hundreds of extensions we've created.
- Maximize data efficiency without sacrificing quality.
- Gain better control and ownership of your data.
- Manage threats and new requirements from changing compliance and data privacy regulations.