WHITE PAPER

The Data Tiering Playbook: A Value-Based Approach to Data Access

splunk>
turn data into doing®

If your organization is like most, chances are you have more data than you know what to do with. Chances also are that you have a significant amount of data hidden, stored in the wrong place or just plain inaccessible, providing little to no insight or value to your business.

To create the most value from your data, you'll have to create a data strategy. But where do you begin?

As a prerequisite to any iterative data strategy, you first have to understand what data you have and determine what's most important to your business — a process known as data tiering. At its core, data tiering is the practice of prioritizing and ranking enterprise data from most critical to least frequently used — accounting for geographical, compliance or format restrictions — and categorizing that data appropriately in a cloud data platform or storage system.

Your greater data access strategy is aimed at allocating data in various hardware classes that can be broadly visible in search efforts, essentially providing you with a framework for routing all of your information. Ideally, this framework can also generate efficient and sustainable data processes that will help you optimize costs, boost ROI and, ultimately, **extract business value**.

Much like a colossal spring cleaning project, the act of sorting through volumes of unorganized data can often be an intimidating — if not overwhelming — effort, creating a paralysis that prevents you from taking first steps. Compounding this paralysis is the need to manage costs, often leading to a patchwork of point tools and solutions, which create blindspots and prevent organizations from fully realizing their data's value.

To ease some of the confusion, this paper provides a methodology for getting started on a data prioritization initiative while articulating the value you can realize from data tiering.

Benefits and challenges of data tiering

For many organizations, the practice of data tiering is becoming increasingly necessary. For one, it creates a sustainable and manageable way to understand where all of your data lives, while giving you the ability to access it quickly from anywhere, when you want it. This level of **data access**, in turn, generates new operational efficiencies — offering better visibility into your environment, boosting monitoring efforts and creating a much clearer picture of your entire data landscape for regulatory compliance auditors, as well as your enterprise.

Data tiering is a critical first step in developing a comprehensive data strategy, entailing three main components:

1. Organizing and prioritizing data for ingestion, according to its value to the business (i.e. use case).
2. Facilitating search capabilities that accelerate time to investigation.
3. Creating a means of cost-effective storage.

But what makes data tiering worth the effort? Cost optimization and the ability to capture more data that helps you to:

- Find a cost-effective solution for each data type and use case.
- Free up an already tight budget to capture deprioritized data.
- Align it to architectures that lean into economies of scale.

While there are many benefits, organizations often encounter barriers that prevent them from taking first steps into their journey. Some of the challenges include staffing shortages and lack of advanced professionals with expertise in data sciences.

Other challenges include lack of time to implement a data strategy that can keep pace with its exponential rise, as well as a lack of overall bandwidth. Because many organizations of all sizes already feel overwhelmed by their data, it's hardly surprising that they also fear that taking on a massive categorization exercise will require them to dedicate more time, technology, resources and skills without any real clarity on the benefits they might later realize.

Other barriers to entry include budget and resource limitations due to competing organizational priorities. Lack of resources can also extend to deficiencies in technology that automate tiering functions. Additional roadblocks include misunderstanding source types and data categories due to a general lack of experience.

Perhaps one of the most significant culprits that impede data tiering efforts is cost. For most organizations, IT Ops is perceived as a cost center by leadership — one that must prove its value by focusing on ways to reduce expenses. When it comes to data tiering, that can mean administrators turn to a spate of disparate point tools to house select datasets in an effort to offset costs elsewhere. The resulting fragmented environment creates even more challenges when the infrastructure and IT stack are so complex and disjointed that cohesive data integration becomes almost impossible to achieve.

However, like any investment, applying more effort into data prioritization up front will ease your workload in the future and set you up for success. Here's how you can take the first steps in prioritizing your data and prove its value to your organization.

Defining your data tiers

Sizing your data into value categories is an important aspect of capacity planning. We've identified three tiers of data based on its value to the business, and determined by frequency of use:

- **Tier 1:** High value
- **Tier 2:** Medium value
- **Tier 3:** Low value

Tier 1 data is considered the most critical and high value data that will be most frequently searched, and what you'll need at your fingertips for monitoring and conducting a breadth of investigations. While it differs for every organization, security and threat data (data needed for detecting and alerting incidents), along with business-sensitive IT operations data (data that helps you trigger a necessary workflow), all could conceivably fall into this category. That said, there is plenty of security and IT operations data that would not necessarily fall into the Tier 1 category. Data such as full-fidelity VPN logs, non-critical IoT data, various access patterns and resource utilization will all likely be better suited for a lower tier.

Tier 2 data will likely include information that is important and needs to be monitored, but is less frequently searched. Supplementary data for low-frequency correlation, forensic investigations, as well as accounting, HR and other departmental applications, training data, internal order processing and other analysis data all could potentially fall into this category.

Tier 3 data is considered of low value to the business because its frequency of use is lower than other data tiers, but it still needs to be stored for historical or archival requirements. Most audit and compliance data falls within this category, which includes information that needs to be kept for regulations such as GDPR, PCI, SOX and CCPA.

Keep in mind that data value may vary across teams and their respective use cases, and thus will be tiered differently. What might be considered Tier 1 to the security or DevOps teams, might be put in the Tier 2 or Tier 3 category by IT operations or other departments. To mitigate these alignment challenges, a data tiering initiative will be most effective when all stakeholders are on the same page with the broader data strategy.

To break down some of these silos, you'll need to bring stakeholders together to gain clarity and transparency about what each team and department are working on and how they perceive the importance of relevant data to their job functions. Here are a few best practices to keep in mind:

- Determine the **use case** they want to address and the business needs driving their tasks.
- Understand the various **data sources**: who owns them and where they live today, and identify the key stakeholders who may be involved and benefit from accessing them tomorrow. Try to truly understand a day in the life of the administrator or analyst, and what the data means to them.
- Regularly **communicate** with all key stakeholders who may be impacted by or use the same/similar data set. Conversely, also understand what data they prioritize first that might be of lower priority to you.

Whether you are tiering your data or collaborating on a more comprehensive strategy, no significant data endeavor can be executed alone. You will regularly need the help of numerous teams and stakeholders across the organization as your data strategy scales and evolves. Opening up channels of communication from the start builds trust, creates transparency and sustains a more pleasant work environment all around.

Implementing data tiering

Now that we have defined the data categories according to their usefulness or value to your business, let's explore how to actually tier the data. A successful data tiering strategy aligns to a **prioritized set of business objectives**, then delineates the use cases, as well as the data sources that drive those objectives.

When it comes to prioritization, you will need to know what data is being generated and is available in your organization. All too often, stakeholders don't know what data exists in their organization, and if they do, they're not sure who owns it or who has access to it. To address these critical knowledge gaps, you can bake in change control request processes into your overall program timeline. Asking the following questions to the right stakeholders will help mitigate delays in execution when it comes time to move the data around.

- How will I access this data?
- Who is using this data and why is it important to them?

Next, you will need to delineate your data by purpose (e.g., "IT infrastructure data"), application or resource type (e.g., "Windows data"), or by sourcetype — log emitter data within an application or system (e.g. Windows CPU performance), unique by topic and data shape. This also will mean understanding its specific use case, which will require asking questions such as:

- What is the importance of this data to my organization?
- What is the need for this data for certain functions and workflows?
- What kind of action will the data require? (i.e., Will it be used for investigation purposes or simply need to be monitored?)

Then, you can fine tune your prioritization exercise more granularly on **use cases** the data would serve, aligned to its corresponding business needs. The following are some of the questions you can also consider asking:

- What is the frequency of need for the data?
- Where does this data live?
- How quickly will I need the data to return a search?
- What are the number of daily, monthly, quarterly or yearly search requirements?
- How do I search for it when the time comes?
- How much does this data cost? What is the estimated value to volume ratio?

To further drill down, you will also want to determine if your most pressing data need is related to stack modernization and architecture, or if it's to enable more of your users with access to a wider swath of data. To support your decision making, you should also be asking the following questions:

- What are the most frequently accessed source types?
- What is the required number of daily searches on a given source type?
- What is the actual value to volume ratio?
- What tools will be relevant to this data?
- Does this data address a multi-use or single-use case?

There are also numerous categories to think about when designing a data tiering strategy as a function of sourcetype, some of which may include:

- **On-prem vs. cloud migration friendliness:** Knowing where you are in your cloud journey is critical. While getting data in is relatively cheap, expenses can add up quickly if you want to move it between clouds and other environments. Prioritize which data can be partially migrated to the cloud with the least amount of effort. Also, while it entails many benefits, supporting a hybrid or multicloud strategy can also cause data silos and tool sprawl, which you'll also need to address to ensure full visibility across your environment.
- **Primary data category:** Each piece of data may be used for multiple purposes, but largely falls into one of a few category types. Selecting the one with the most common usage will help align better for cost purposes.
- **Highest use case category:** The access frequency indicates the level of the data's value, as well as the costs you're willing to incur for storing it. Understand that some data moves through a full lifecycle of monitoring to compliance over time, and should be treated differently than pure compliance data.
- **Regional data restrictions:** For data originating in a different region, there may be regional compliance requirements, such as personally identifiable information (PII) masking or personal data that needs to remain within those borders. Understanding those restrictions can ensure you build a routing, storage, and search architecture that fits your business's needs.
- **Fidelity requirements:** Fidelity requirements are often tied to the use case category, and help determine the type of data preparation that's required, as well as the optimized storage and search tools that are needed.
- **Volume reduction/data preparation:** Knowing if the data needs to be transformed or prepped in any way helps determine the corresponding data movement tool, if there is one.
- **Data source:** The data source may influence the agent and data collection strategy.

As a final step, you will need to map your categorizations to the right storage tiers, architectural approaches, data preparation and formatting needs. Again, mapping is highly contingent on your own unique data needs, business strategy and budget. Consider the following questions when making this judgment:

- Which solution maps best to your organization's larger data strategy?
- Which solution reduces the number of tools involved in the process?
- Which solution will cost less?
- Which solution will provide longer-term success rather than a short-term win?

Categorizing your data up front isn't always a simple process — throughout all phases of your value alignment exercise, you will need to continuously and iteratively evaluate use cases, performance and the level of investigation. But once you have classified your data accurately and to your satisfaction, you will have laid a solid foundation for the steps ahead.

Realizing value from data tiers

Prioritizing your data is vital to reevaluating your organization's infrastructure as data volumes increase in age and value. As such, it is important to take a hard look at your environment to determine what is working optimally and what can be made more efficient. Reducing the overall volume of your data through tiering will free up compute cycles that will drive down storage costs and help facilitate more accurate searches. Improving search efficiency, in turn,

will create additional cost savings by allowing you to quickly address outages and incidents, while putting the accelerator on the audit process.

As part of this evaluation, you might also want to consider a governance approach, which will establish the best processes, roles, policies, standards and metrics that will enable effective and efficient use of your data, ensure its quality and security across your organization, and allow you to achieve your business goals. That will mean determining who in the organization is accountable for managing and monitoring the data tiering strategy (some organizations have a chief data officer, who becomes the de facto owner within the office of the CIO.) They will also be responsible for determining the frequency with which they review the strategy and the decisions they make around the change control process.

Additionally, you will also want to strongly evaluate the number and type of tools in your data arsenal. As previously mentioned, organizations often resort to using a variety of piecemealed solutions or storage for lower data tiers that might provide some short-term savings. However, managing numerous tools that are not communicating often results in data that is hidden, not easily accessed or simply unknown — all of which can have a serious and costly impact to your business over the long run. What cost savings you might realize initially will be overshadowed by the need to add more staff, expertise and overall visibility to compensate for the non-integrated tools. Consequently, competing on price and cost becomes a losing game.

For mission-critical tasks, mismatched and siloed tools can also have dire consequences. In the event of a breach, for example, you will need to determine where the attackers are in your network, how long they have been behind the firewall, where they originated and where they are going next to understand the magnitude of the threat. To get that 360-degree view, you will need network and proxy data, audit and compliance logs, performance metrics and compute utilization. In short, you will likely need access to all three tiers of data at any given time in order to fully realize the scope of the attack, identify and contain the threat. Trying to find that same threat information across multiple systems and tools, however, might prolong the time attackers spend in your network, or result in inaccurate investigations, while adding elevated costs to the price of cleanup and remediation in the aftermath.

This example highlights the value of **federated search**, which allows organizations to correlate and search across all three tiers at once using a centralized platform custom-built for querying machine data. For organizations, that means a holistic view across an entire data ecosystem — even as it remains in its silos — translating directly into faster time to detection and resolution.

Of its many benefits, federated search will help you:

- Filter and route data efficiently based on its value tier
- Consume both low-yield and high-value data cost effectively
- Store data in ways that allow you to minimize costs and maximize value
- Access your data via search, regardless of where it is stored

The biggest value derived from data tiering is the ability to achieve **rapid time to investigation and action** — accelerating your search so that you can get more done in less time. The ability to access information quickly is especially critical in light of compliance and other regulatory mandates, when prolonged searches can also incur hefty financial and reputational penalties that can quickly add up and impact your bottom line.

Looking ahead, data tiering will continue to be a pivotal first step in a comprehensive data access strategy that will pave the way for all of your search and storage efforts. Ultimately, the longest data journey starts with the first step. By taking initial steps to understand and prioritize your data, you will start the process of breaking down information silos and opening visibility and channels of communication, while creating new efficiencies, increasing ROI and generating long-term value that will benefit your organization for years to come.

Download the “The Essential Guide to Data” here.

22-25319-Splunk-The Data Tiering Playbook-WP-103