EDGE DELTA

Optimizing Observability

A Guide to Minimizing Logging Costs and Enhancing Performance

Many log monitoring solutions were originally architected when ingest volumes of data were measured in GBs per day.

Observability Practices Need a New Way to Solve Data Gravity

Data gravity refers to the difficulty of moving data as it grows in size and volume. As a byproduct, deriving value from your data becomes cost prohibitive and slow.

Data gravity is affecting observability practices in particular because log data volumes have ballooned over the past several years. At the same time, the way that we derive value from log data has largely stayed the same. Observability and monitoring tools have been around for 30+ years. In the early 2000s, the industry shifted towards a centralized approach, in which customers aggregate their data in a platform before they analyze it.

As Gartner® explains, "Many log monitoring solutions were originally architected when ingest volumes of data were measured in GBs per day." They continue, "Today, many large enterprises are ingesting higher than 10TBs per day, with an increasing number observed in the 100TB range.[1]"

[1] Gartner, *Cool Vendors in Observability and Monitoring for Logging and Containers*, Published 27 April 2022
GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Increase in Log File Consumption by Large Enterprises

Log Volume Ingested

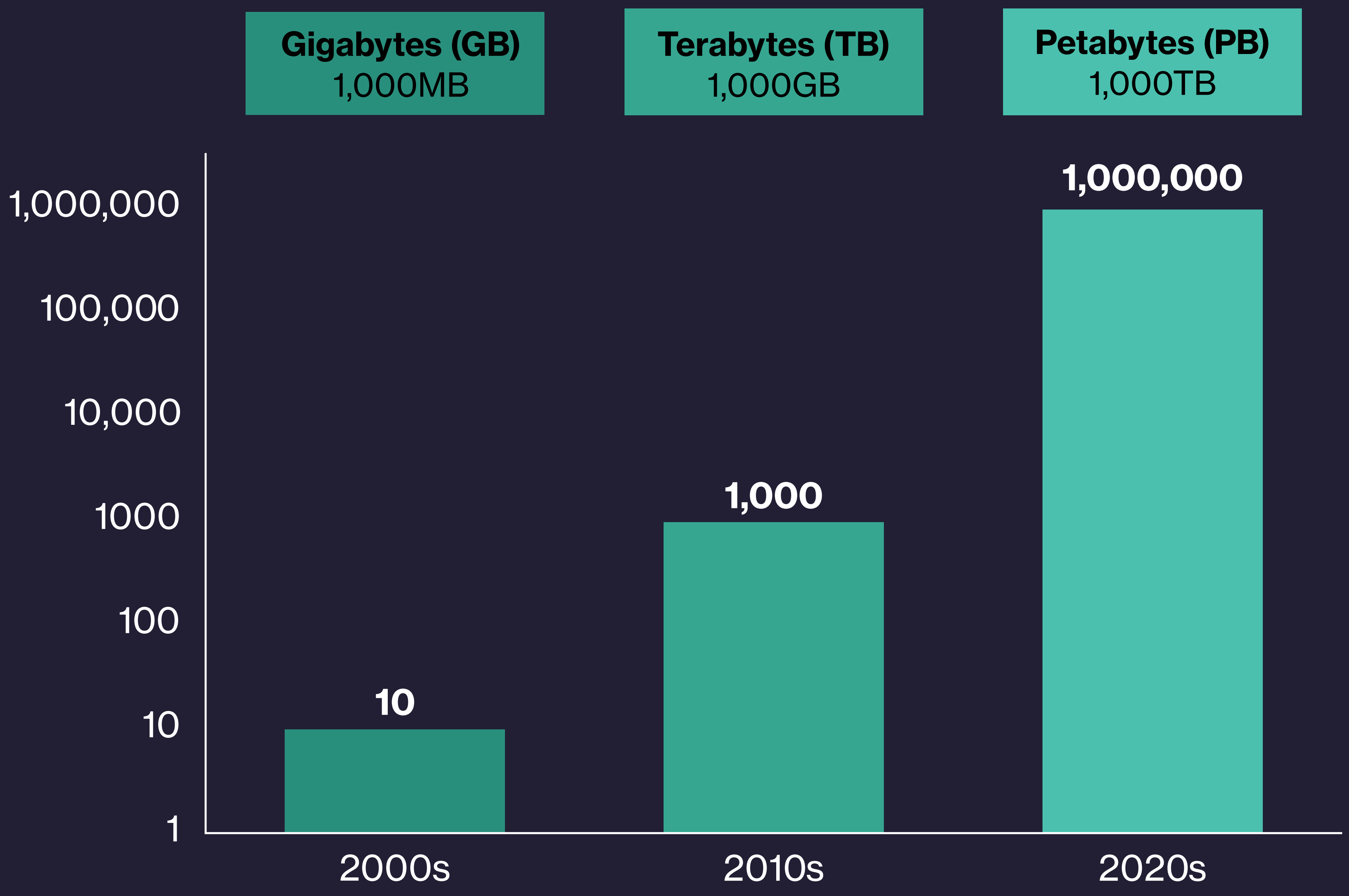


FIGURE 1: *Cool Vendors in Observability and Monitoring for Logging and Containers*, April 2022

The combination of unbounded data growth and outdated observability and monitoring architectures creates several challenges, ranging from poor performance to excessive costs.

In this guide, we'll offer recommendations on how customers can mitigate data gravity. We'll help you balance the needs of your team (e.g., uncovering real-time insights) and the business as a whole (e.g., adopting a budget-friendly solution). But first, let's look at the challenges in greater depth.

Data Growth Makes Observability Costly

In the previous section, we established that most observability and monitoring platforms were architected in a previous era – when data volumes were comparatively smaller. These platforms typically charge customers based on the volume of data they ingest.

As a byproduct, it's becoming prohibitively expensive for companies to analyze all of their log data. How dire has this challenge become? In Datadog's 2023 first-quarter earnings call, it was revealed one cryptocurrency company spent a whopping $65 million on observability the previous year.[2] While this is an extreme example, there is growing sentiment that the value customers receive from their observability tools doesn't match the cost they pay.

[2] Datadog (DDOG) Q1 2023 Earnings Call Transcript

Reducing Observability Costs

In order to reduce costs, organizations have tried several strategies to limit the volume of log data they ingest into their observability platforms.

Dropping Events	100% of this data type is discarded
Sampling Events	1 out of every N of this data type is delivered, the rest are discarded
Dynamic Sampling	Low-volume data of this data type is delivered at 100%. As volume increases, the percentage of dropped data increases.
Suppression	No more than N copies of this data type are delivered per unit of time.
Parsing + Trimming Events	Removing unnecessary, unwanted, or overly verbose parts of an event.

These strategies can help companies reduce log ingestion volumes in the short term, but there is still an underlying issue: data volumes continue to grow and legacy architectures can't handle this scale.

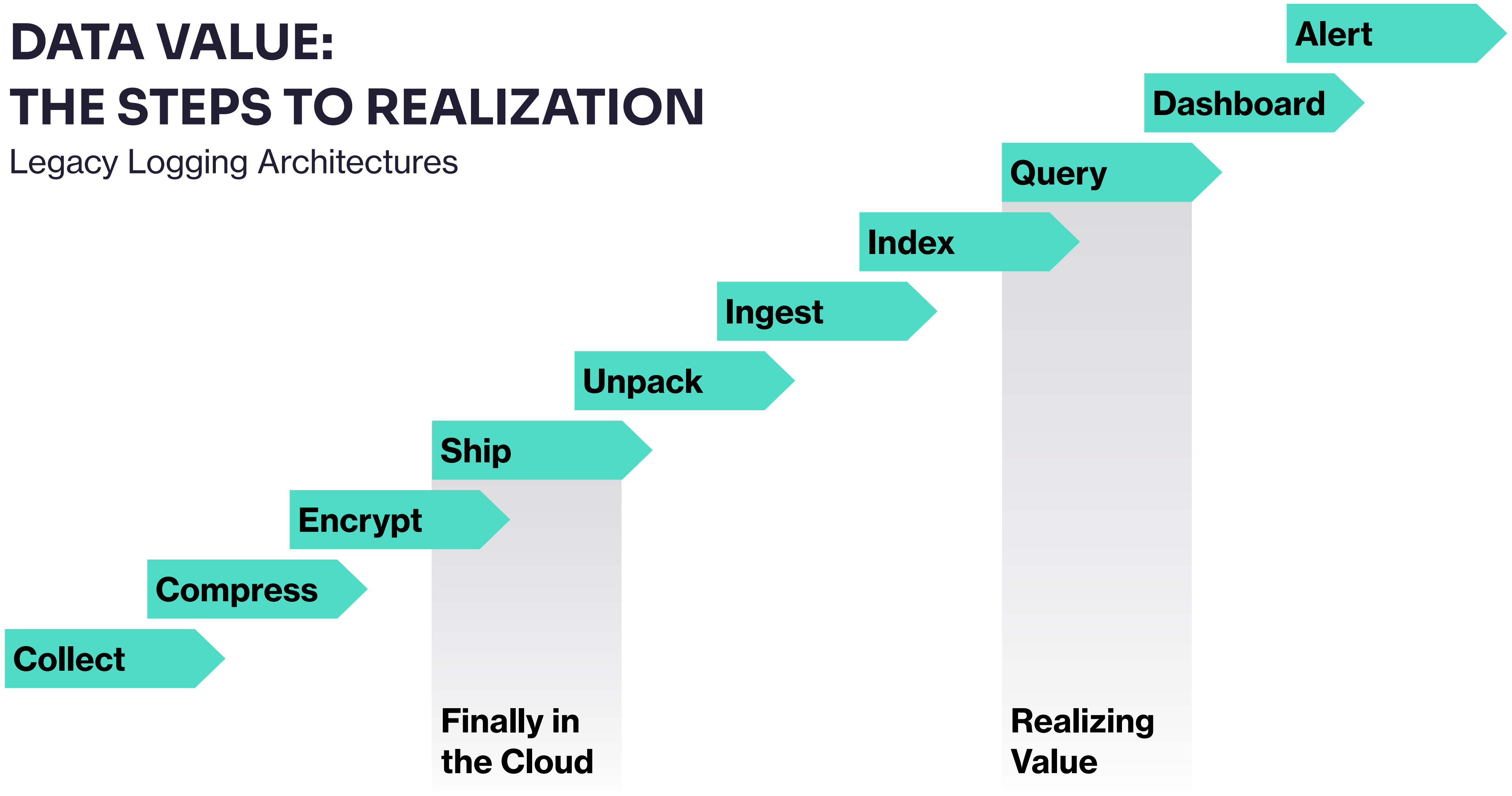
Additionally, these approaches require the exclusion of data from your observability platform. This requires you to foresee which data you'll need and which you are safe to neglect – a difficult task that creates blindspots for your team no matter how accurate you are. Blindspots will only get bigger as data volumes grow.

Legacy Architectures Degrade Performance

The primary challenge organizations face is related to the cost of their observability tooling, but there is also the issue of performance.

DATA VALUE: THE STEPS TO REALIZATION

Legacy Logging Architectures



Centralizing log data is a complex process with many different failure points and bottlenecks that can slow analysis. Once data is finally indexed, your platform has to batch-process massive data volumes to generate analytics and return queries.

This process worked when data volumes were smaller, but now it creates bottlenecks that increase in severity as data volumes grow. At best, you can experience a delta between when your data is created and when it is finally analyzed. At worst, one of these points can fail and inhibit analysis altogether. Either way, teams struggle to support real-time use cases with this model.

Balancing User and Business Needs

Based on customer pain points, it's clear that the underlying challenge is architectural. It is not sustainable to collect, ship, ingest and analyze TBs of data each day. Before jumping to a better observability architecture, it's important to first audit how teams are using their data. There are typically two use cases:

REAL-TIME ANALYTICS

Running scheduled queries, populating dashboards, and triggering alerts to support real-time monitoring.

AD-HOC QUERIES

User-generated queries to troubleshoot production issues, investigate security incidents, and answer other questions.

Vendors have tried supporting both use cases with a centralize-then-analyze approach, but it's possible to support each use case by decoupling where you analyze data from where it's stored:

+ Process data upstream to support real-time use cases

+ Route raw data to a low-cost storage target for ad hoc queries

This gives teams the real-time insights they need from their data while protecting the business from excessive costs. In the sections that follow, we'll explain how this approach works at a deeper level.

Stream Processing for Real-Time Analytics

Many teams rely on software agents to collect and route data, and centralized pipelines to pre-process data before it's ingested in an observability platform. However, teams can use similar technologies to push compute upstream to where data is collected. This enables you to derive analytics to better support real-time use cases. Stream processing your log data as it's created allows you to populate monitoring dashboards with lightweight analytics versus complete raw datasets.

This approach provides several advantages. First, teams no longer need to wait for data to be indexed before they are deriving value – they can analyze data and trigger alerts faster than before.

Second, the analytics created are based on 100% of your data rather than the percentage of data that is available after filtering and sampling rules have been applied. You'll receiving greater data coverage and able to monitor a broader set of services and systems than before.

Lastly, teams can dramatically reduce costs, given they are indexing analytics. In many cases, teams can avoid indexing the underlying raw data.

TAKEAWAYS

+ Derive analytics without paying to index complete raw datasets

+ Reduce the total volume of data you ship downstream to optimize cost and performance

+ Deliver faster insights to support real-time use cases by processing data upstream

Low-Cost Storage for Ad-Hoc Queries

Deriving insights to support real-time use cases is undoubtedly time-sensitive, but ad-hoc queries are typically not as urgent for your team. Especially for datasets that your team queries less frequently. This is an area where you can prioritize other factors over performance, including:

- Cost-effectiveness
- Data coverage
- Data residency and security

It makes sense to find a provider that enables you to move data to more than one destination. In doing so, you can identify a cost-effective alternative to store lower-priority data while routing higher-priority data to your log index.

This approach enables you to right-size costs because you're allocating your observability license to your most frequently accessed data. In addition, data coverage is increased, so teams won't have blind spots.

Lastly, this approach provides greater control over where your data resides. Perhaps you need to comply with regulatory requirements, or your security team is more comfortable with data staying within your environment. In this situation, you can rely on a customer-owned data store to more easily meet these needs versus using a pure vendor-hosted offering.

TAKEAWAYS

- Right-size costs by allocating frequently searched data to hot storage tiers, and lower-priority data to cost-effective storage
- Greater data coverage: no more “blindspots”
- Gain greater control over data security and residency to meet compliance and security needs

A New Approach to Observability

As a byproduct of log data growth, the current model of centralizing all raw data is no longer sustainable. Instead, teams need to audit their log data use and adopt a new approach that balances individual and business needs.

When you take this approach, you may land on the options detailed in this guide. You may also find alternatives that meet your team's unique needs. For example, perhaps you begin stream processing your log data for real-time analytics but continue relying on an index to maximize querying speed.

No matter the outcome, you need open and modular tooling to maximize the value of your log data both today and in the long term.

About Edge Delta

Edge Delta is a new way to do observability. We process your data as it's created and give you the freedom to route it anywhere. Make observability costs predictable, surface the most useful insights, and shape your data however you need.