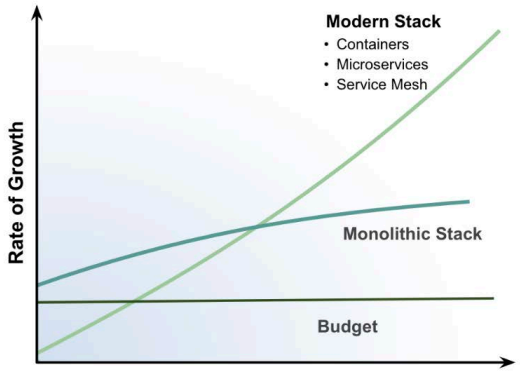
Event Data at Scale: The Foundation of a Modern Business

Are you paying too much to collect, keep and use your data?

Real-time event data is the most accurate and granular view of the health and performance of a digital business. Because of cost, complexity, and scale challenges, most organizations believe it is impossible to collect, retain, and use all this data.

Scalyr believes it's essential.

Data is the lifeblood of an organization. For engineering, product, and operations teams, the real-time performance data of a system is essential for problem detection, isolation and correction, incident management, forensics, planning and optimization, trend analysis, behavior insights, and more. It feeds a variety of use cases as well as AI and ML engines. In addition to being a reactive tool, the use of this data can help improve performance, create competitive advantage, increase user satisfaction and engagement, and catalyze new product innovation.



In modern stacks, the data that is generated increases exponentially with the move to microservices, containers and the cloud. Budgets are not keeping pace.

Recognizing the operational importance of this data, there are more than 25 years of history of solutions meant to collect and organize the data for analysis. Vendors have brought to market software for on-site hardware, and more recently cloud-based solutions. But the sheer growth and quantity of data have outstripped the capability of these solutions to affordably scale as needed. In response, the vendors have come up with workarounds including sampling, indexing, increasing cluster size, and tiered storage to cope. Metrics were thought to be a solution, discarding context in favor of a materialized calculation. But these approaches have also become complex, expensive, and discarding data may impede problem-solving.

SCALYR™

There has to be a better way.

Integrated Event Datastore

Scalyr believes that every company, regardless of scale, should be able to affordably collect, keep and analyze all its event data without sacrificing context or performance. Organizations should not be forced to decide today what data will be valuable in the future. Nor should organizations be forced to discard data and critical context, because of affordability concerns.

Industry-Leading Breakthrough

Your data is only valuable if you can afford to collect, retain and use it. Current practice has been to harvest or summarize data, and to store what is left in cheap storage at the cost of availability. With Scalyr, these tradeoffs are unnecessary. Scalyr customers can log 200 terabytes per day, retain everything for as long as needed, and quickly analyze everything that is stored, affordably.

Scalyr offers a breakthrough in scale and affordability. Scalyr compactly retains data in S3 with metadata enforcing segregation of customer data, fast retrieval into Scalyr's query machines, and efficient use of storage. Scalyr's customers benefit from the industry's lowest cost of ownership, especially when compared to do-it-yourself solutions.

Scalyr's multi-tenant cloud architecture makes event data useful by putting massive scale at your fingertips without sacrificing query response time. Scalyr separates storage from query services and allocates every CPU against each query as it's received. This is accomplished by having stateless services scale horizontally and independently within Scalyr's architecture. An incoming query is distributed across the full multi-tenant cluster in parallel to deliver extraordinary compute power.

Scalyr makes it affordable to collect, retain, and analyze all of your application and infrastructure operational data.

Scalyr's Breakthrough Technology

Scalyr's Approach	versus Competitive Solutions
Keep all of your data active and searchable. All event data is "hot" and ready for analysis.	Store data in tiers and pack data into "cold" storage to minimize costs. "Rehydrate" data when it is needed.
Capture, retain and be able to immediately analyze all of your event data.	Index the data before availability, maintain a small amount of data for analysis and archive the rest.
Separate storage from compute, and allocate every CPU against each query.	Dedicate shards and allocate storage for a given customer. Query response time depends upon allocated vCPUs.

Breaking Through Sound Barriers

Is it possible to collect and keep all of your event data, in one place, so that you can search it, use it, and get value from it? Three things are needed, working together: scale, affordability, and performance. Traditional solutions have unintentionally created barriers for each.

To keep and store data, the underlying system must scale, be agnostic to data type or structure, and be accessible through APIs or other automated agents to tap and use the data.

Scale: If you're going to collect and keep all your data then an integrated event datastore is required that can accept data from any source, structured or unstructured, and can scale up to meet business needs.

- Scalyr collects logs, metrics, traces, alerts or any of the dozens of other events needed to provide a holistic system view.
- Scalyr users can ingest up to 200 terabytes of data per day, or more, without indexing delays. And you can keep the data as long as you need it.

SCALYR

Affordability and **performance** are integral to scale. If you can't afford to collect and keep the data, then scale and performance are meaningless. If you have the data, but it's too slow and painful to use it, then the data is essentially useless.

Affordability: Affordability is critical. If organizations can't afford to keep the data for as long as they would like or keep it at all, then context and insights are lost forever. Costs to collect and retain data should be low enough that throwing data away is not a viable consideration.

- Scalyr has achieved an order of magnitude cost advantage over competitive solutions.
- Data can be retained long term at S3 storage rates.
- Scalyr's pricing rivals do-it-yourself options using open source solutions, and you get a full service, full-featured SaaS solution out of the box, with no tuning or maintenance required.

Performance: For most companies, queries may take minutes or longer to return a result, disrupting workflow and thought processes. If a web page takes more than three seconds to load, most people will abandon the effort. Users become frustrated or simply avoid using the system and tapping the data. We live in an always-on world of instant gratification. You move at the speed of thought, and so should your analysis tools.

- Over 90 percent of Scalyr searches return in less than a second.
- Customers who have done head to head comparisons of Scalyr vs. Splunk report Scalyr delivers 10x-60x faster query response times.

How Does Scalyr Achieve This Breakthrough?

Scalyr offers customers up to 200 terabytes per day of log ingestion or more without sacrificing fast and affordable search and backed by long-term storage. Scalyr's 200TB/day service surpasses the previous industry record for cloud-based services by 50%. How can this be possible?

Scalyr was the first to pioneer a non-indexed solution with a massively-multi tenant compute architecture and purpose-built columnar data store. Scalyr has now expanded its design with autonomous, stateless services for ingestion, search, metric generation, parsing, analysis, event reporting, dashboards, and data export. These services are decoupled instances, allowing the Scalyr architecture to deliver unprecedented scale, affordability, and performance. Scalyr's log management functions scale-out on demand, supporting continuous capacity expansion.

The result is a network-effect built deep into the foundation of the architecture. **The more data Scalyr ingests and the more customers using the service, the faster and the more affordable the system is for all customers.** Scalyr gets faster and more affordable as usage grows, the opposite of all other log analytics systems.

An example of the power of this architecture is Scalyr's ability to apply all of its cloud computing resources to customers' queries delivering search results almost instantly across terabytes of data. Even complex queries including joins and pipes return values within seconds.

Scalyr's architecture is in sharp contrast to the approaches used by most alternatives.

Scalyr's Industry-Leading Breakthrough

Area	Scalyr	Competitive Contrasts
Scale	Scalyr's cloud-based modular architecture consists of stateless services that can scale out autonomously while maintaining consistent features and performance.	Competitors have built fragile and fixed-configuration architectures that interlock storage and compute and create dependencies that cap scalability.
Data Retention	Scalyr's effective use of S3 allows for long-term affordable event data retention and simultaneous delivery of fast log analytics.	Competitive solutions require painful decisions and trade-offs between data retention, availability, and cost.
Data Availability	With Scalyr, all data is hot. There are no compromises between price and availability. Logs available and searchable immediately after input.	Alternative solutions will delay the availability of data for indexing and later ask you to move data to cold storage to save money. Cold data needs to be "re-hydrated".
High-Cardinality Data	Scalyr is designed for highly cardinal data. Examples: transaction IDs, customers #'s, location information, different types of rates and measures cannot be indexed as key-words.	Competitive solutions are based entirely on keyword indexing, which was designed for business information systems, not primary operations data.
Search Speed	Scalyr uses stateless query machines that fan out to deliver results within a second. Search for any data or value instantly.	Competitors have designed stateful search algorithms that use an index to find keywords and create lists. Indexing and searching data can take hours.
Multi-Tenant Cloud	Scalyr uses multi-tenancy to gain economies of scale and horizontally distributes stateless compute agents across the entire cluster to deliver breakthrough ingest data rates and query response time.	Competitive single-tenant solutions offer little scale economy. Their 'lift-and-shift' model offers fewer advantages in the cloud than Scalyr's approach.
Compute and Storage Segregation	Scalyr compactly retains data in S3 with metadata enabling segregation of customers data, fast retrieval into query machines, and efficient use of storage.	Competitors create stand-alone resources, just for specific customers, limiting the efficiencies of scale that can be gained from a cloud-based service.

Optimized for High-Cardinality Data and Its Use Cases

Keyword indexing refers to a method for structuring and organizing data so that it can be easily searched. Keyword indexing is optimal for many use cases and was designed to help people filter through massive amounts of information to find a subset of "the best" examples. When you search the world wide web using a keyword and get back the top websites that represent your topic, you've made excellent use of keyword indexing. It is optimized for human legible terms, relatively-static content, like web pages and low-cardinality and low-dimensionality data sets. https://www.scalyr.com/blog/keyword-indexing-is-flawed.

Event data are machine-generated and differ in two main ways:

- High-cardinality (a high level of uniqueness)
- High-dimensionality (many variables)

High-Cardinality: This refers to the level of data-uniqueness. Event data is equivalent to application exhaust. It's machine data and generally has a very high level of unique identifiers, like session IDs, IP addresses, etc.

High-Dimensionality: This refers to the number of attributes in a data set. With event data, not only do there tend to be a high number of dimensions but making sense of the data may mean using, filtering and/or combining different dimensions to answer a question.

SCALYR

Engineering use cases for operational data, such as logs and metrics, are the opposite. A sample of errors, failures, or breaches isn't helpful. You don't want to know the top ten times a server failed, or some examples of a 404 error, or a few times a malicious user entered your system, you need all the incidents, and you need to rapidly visualize, organize, and filter the information so that you can dig into the root cause and see patterns.

Indexes provide flat results: a list. Users of operational data are often trying to solve a problem. As such, the system needs to zoom out to visualize the aggregate data, zoom down to see specific details, pivot around unusual results, and zoom back out to see the big picture again. Over and over. Quickly. The index was never designed for this type of workflow.

Conclusion

Primary operations data is the lifeblood of managing a digital system, service or business. Until now it was difficult to keep and analyze the critical data emitted by these applications and infrastructure, especially as they migrate to a modern stack.

Scalyr believes it is not impossible, it is essential, to collect and retain all event data from applications, services, and infrastructure. Scalyr's revolutionary multi-tenant architecture separates storage from compute, and allocates every CPU against each query as it's received, delivering a breakthrough in scale, affordability, and performance.