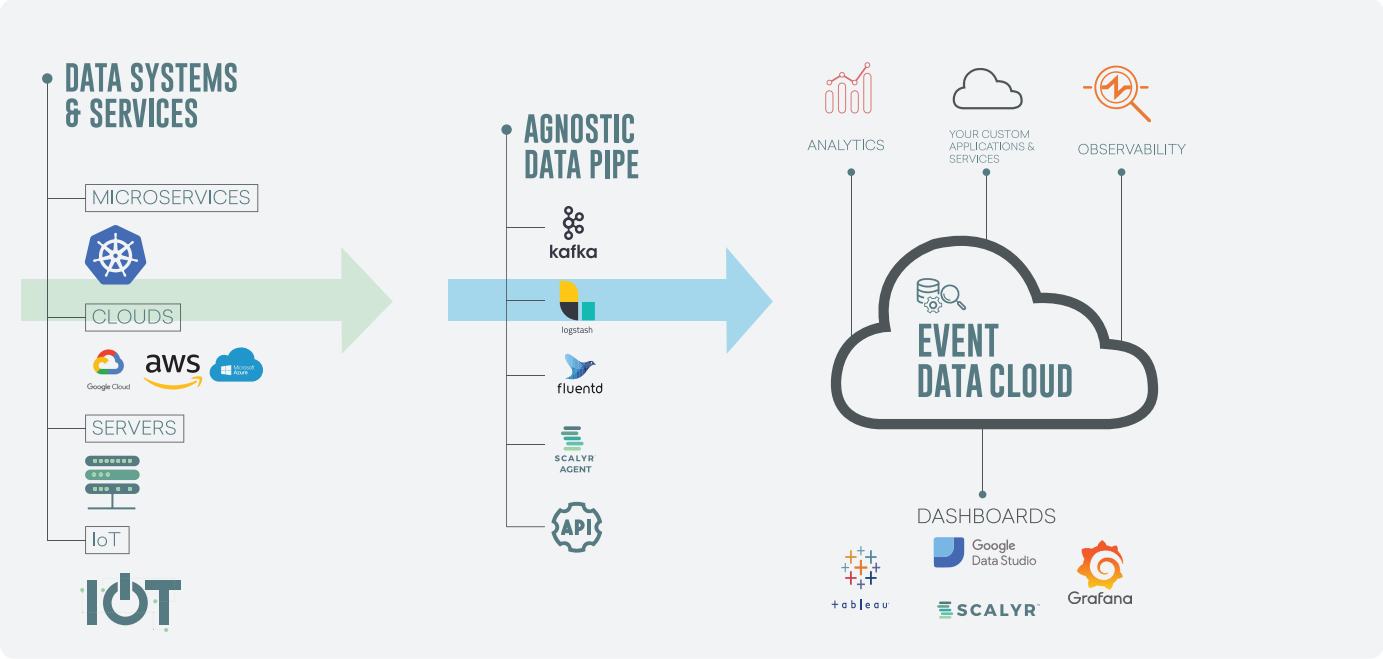
EVENT DATA CLOUD

The purpose of an event data cloud is to unleash the full potential of event data for a range of applications and analytics for use across engineering, IT, product planning, and business planning. It is where companies catalog/organize/optimize data for analytics and observability use cases. The event data cloud provides the most current, accurate and complete view of a digital system/service and serves as a platform for applications and users that need access to this data.



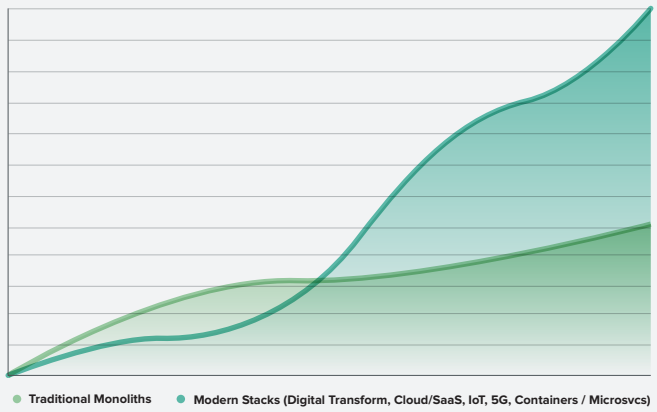
Real time event data is the most fundamental and granular view into the health and performance of digital systems and services, as well as insight into security issues, performance trends, usage patterns, user needs, optimization, and so forth. The data is rarely exploited for these applications today because of cost and complexity of ingesting, retaining and analyzing this data. The event data cloud addresses these problems while enabling new use cases to accelerate growth, improve availability and reliability, and increase insight and intelligence.

This paper will discuss the problems, adjacent categories, first and second generation solutions, technical requirements, use cases and business benefits of the event data cloud.

THE MAGNITUDE OF THE PROBLEM

As companies and entire industries digitize and digitally transform, the engineering teams that design, create, and optimize the digital service become the center of the organization. Event data is the only mechanism for these teams to reason on a digital system, understand its complex behavior, and expose its secrets. Unleashed, event data can be used across the product and planning organization to analyze trends and indicate how system performance and business outcomes are correlated, and by IT teams to support security and compliance requirements.

EVENT / MACHINE DATA AS GENERATED BY MONOLITHS VS. MODERN STACK



The volume and importance of event data grows exponentially with the move from monolithic to modern stacks, driven by the use of cloud, containers, and microservices. The amount of data generated by the digital system can explode by 10-100x, but budgets to manage and analyze this data do not increase just because architectures modernize.

EVENT DATA CLOUD DEFINED

AN EVENT DATA CLOUD IS OPTIMIZED TO

- economically ingest, retain, and search event and machine data
- enable first and third-party interactive analytics, summation, and visualization of the data
- handle petabyte-scale capacities at low latency for ingest and search

Events have two fundamental properties: Events are time bound, which means each event happens at some exact point of time. And they are data-unbound, which means that each individual event can carry many data fields, and a near-infinite number of data dimensions. As such, the event data cloud has two primary and *interdependent* dimensions: **Time and Data.**

TIME DIMENSIONS

• Event data clouds ingest real time data, with time-based stamps, and use time as a prime partition.

• Latency of data usage is critical. Data must be made available for use within seconds of it being created, and search / usage of the data must be fast enough to support real time use cases like problem isolation and resolution.

• Queries always have a time component to the search; the search is bound by a time range.

DATA DIMENSIONS

• Event data clouds are optimized for structured and unstructured data, and anything in between.

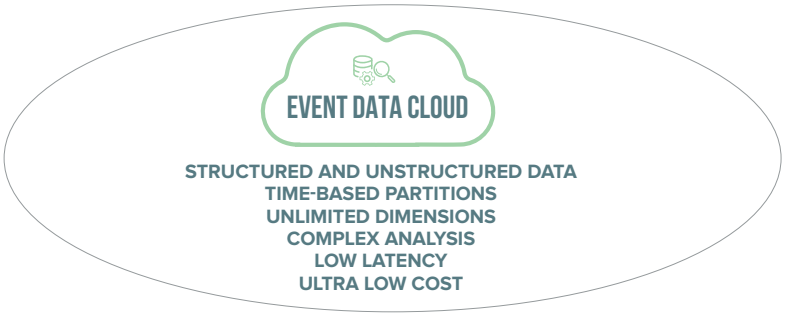
• The data has the properties of high dimensionality (thousands of fields) and high cardinality (unique identifiers and low replication).

• The data has a time-dimension; is time-stamped and time partitioned (as mentioned above)

To reason effectively on event data, the event data cloud needs to provide users the ability to work at the speed of thought. Latency in ingest and slow query speed can impact workflow, particularly for incidence management and response, but also for extracting insights and optimizing performance. A company must be able to afford to collect and retain the data, so low cost, generally measured as $/GB ingested and/or searched, is critical.

FINALLY, EVENT DATA CLOUDS MUST ENABLE SERVICES OF DIVERSE TYPES TO

- send their data to the event data cloud, and have it represented there in a manner that enables high performance analytics; and
- access event data, analytics, and summarizations through APIs so that these services can power specific work flows and visualizations



WHAT AN EVENT DATA CLOUD IS NOT

Now that we understand what an event data cloud is, let's understand what it is not. An event data cloud is similar in many important ways to familiar, adjacent categories, and yet it does not fit neatly into any of them. Understanding the comparative similarities and differences is important.

AN EVENT DATA CLOUD IS NOT THE FOLLOWING

CLOUD DATA WAREHOUSE

An event data cloud is a lot like a cloud data warehouse, in that it stores massive amounts of data in the cloud and serves it up on demand. The cloud data warehouse is optimized for relational business data and complex financial/business analysis using SQL queries. The event data cloud is optimized for time-partitioned, event data, and the associated use cases and users. And, its analysis is fast by comparison, because many of the use cases are real-time in nature.

- The event data cloud is also not a data lake. A data lake is a general term to describe data written to a central file system or an object store, that typically doesn't concern itself with providing answers in real-time. Data lakes are easy to form, but are often expensive to query.

DATABASE

The event data cloud can look a lot like a cloud database, because there is a database inside -- a specialized, highly efficient database optimized for event data. A generic database is built to support a great variety of use cases. Like a SaaS database, the event data store provides programmatic access to a managed service that allows developers to store and access data. It is a deeply integrated collection of databases and services, which enable a seamless and near real time experience for developers to leverage its underlying data via APIs and SDKs. But unlike a database, the event data cloud is not an ACID compliant, nor does it implement SQL or RDBMS structures. Its core is typically a NoSQL columnar store.

ENTERPRISE SEARCH

The event data cloud IS about search; blazing fast, real-time search. But it's not about 'enterprise' search across enterprise files or relatively static web/human legible data, generating top 10 lists, and understanding synonyms. It is search optimized for event and machine data that has the properties of high cardinality and high dimensionality, and is designed to return *complete results* on massive data sets.

STREAMING ANALYTICS

The event data cloud shares properties with streaming analytics. There is extraction, transformation and loading of data (ETL), and real time processing of data. But the event data cloud is not optimized to analyze and shape a user's real time web or service experience and change the content or flow of that experience. It's designed to support complex data, ingest it in real time, and provide visualizations of data and answers to ad-hoc and repetitive queries.

EVENT DATA CLOUD IN CONTEXT OF THE MARKET

The event data cloud fills a market void and provides a solution that is purpose-built for event data and the use cases, often real-time, that use event data. Scale, performance, and cost are all critical to creating an effective and efficient event data cloud.

First and second generation solutions use very different technologies, and thus are able to achieve very different levels of scale, performance and cost. Available solutions, and the underlying technologies, are discussed below

FIRST GENERATION SOLUTIONS

First generation attempts to deliver real time insights into operational data in the pre-cloud era were repurposed enterprise search technologies that use indexes to organize the data and are designed to search the web and enterprise file systems. The index technologies are optimized for relatively static, human-legible information and designed to provide a representative sample of data for any given query (think 'top 10 lists' for web search). These early solutions solved an important problem, and are great for their respective use cases, but often fall short for event data. Rarely does an engineer want a representative sample or the top ten times their system failed. They want ALL the data that fits their query parameters, and they often need to search numerical ranges, wildcards and summarize millions of matches.

There are also widely used open source solutions deployed on-prem, as well as hosted in the cloud, geared for engineering teams. These early solutions equipped companies to centralize, analyze and visualize data, but struggle with scale, complexity, and affordability for event and machine data.

Event data clouds can be built from first generation technologies, but there are trade offs, especially for engineering use cases and users. Many first generation solutions are geared to support IT use cases, and are optimized for the specialized power-user in these departments.

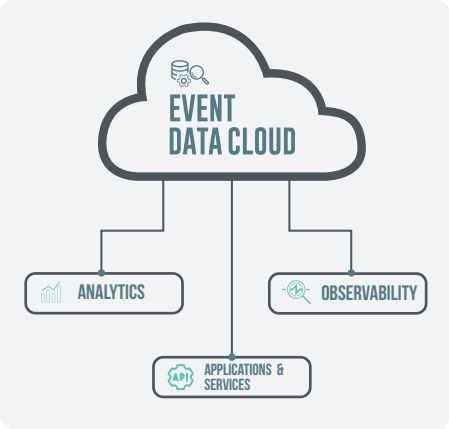
EARLY PROVIDERS INCLUDE BUT ARE NOT LIMITED TO

- OpenSource, DIY and managed Elasticsearch / ELK services.
- Offers built on Elasticsearch (SumoLogic, Datadog and others)
- Splunk

These are successful companies and technologies, well suited to their intended use cases. The growth of these index-based solutions exploded and became the de facto standard over the past two decades. But the Gordian Knot of performance, affordability, and scale for event data have hampered first generation solutions and prevented companies from unleashing its full potential. It is just too expensive, and too cumbersome, to fully exploit this asset for a broad range of use cases with first generation technologies.

The modern event data cloud fills this void, and is purpose-built for engineering use cases, and the optimization of digital systems. If you started on a first generation system for event data, and have challenges with scale, cost, or performance, an event data cloud may be your next act.

THE EVENT DATA CLOUD



Second generation technologies have emerged to deliver scale, performance and affordability. Like their first generation counterparts, these solutions offer the option of open source and vendor-hosted services that support a range of observability, security, and other applications.

The event data cloud ingests petabytes of event data, in real time, without latency. It is cloud native, optimized for microservices and containers, and able to burst to handle major fluctuations in data volumes and usage patterns. It supports automated, repetitive and ad-hoc queries as well as complex analysis. It uses a columnar store or modernized data structure to avoid the limitations and problems of traditional indexes in supporting high cardinality, high dimensionality event data. Just as importantly, it removes price as a barrier that prevents companies from collecting, retaining and using this data for a variety of applications.

EXAMPLES OF MODERN EVENT DATA CLOUD SOLUTIONS INCLUDE THE FOLLOWING:

- Self-hosted, open source solutions: Apache Druid describes itself as “A modern cloud-native, stream-native, analytics database” designed for workflows where fast queries and ingest really matter. Druid excels at instant data visibility, ad-hoc queries, operational analytics, and handling high concurrency. Consider Druid as an open source alternative to traditional data warehouses for a variety of use cases. A self-hosted event data cloud can be built from Druid.
- Vendor-hosted: Scalyr provides an event data cloud, as a service, in two forms.
  - as part of a comprehensive log analytics and time series solution with a full featured UI,
  - as an OEM/API-driven solution that operates ‘under-the-hood’ of other analytics services

TECHNICAL REQUIREMENTS FOR AN EVENT DATA CLOUD

The purpose of the event data cloud is to unleash the full potential of event data, therefore anything that stands in the way of scale, performance and affordability has to be engineered out of the system. Said differently, the event data cloud must be architected to simultaneously deliver massive scale, low cost and fast performance. To do all three together requires a specialized architecture for event data and the associated use cases.

TECHNICAL REQUIREMENTS INCLUDE THE FOLLOWING

Flexible Data Schema	○ Support of structured, unstructured and semi-structured event data.
Separation of storage and compute	○ Independent and horizontal scaling compute and storage ○ Able to handle large and unexpected bursts of data
Columnar store, minimizing or removing undue overhead of indexing	○ Columnar store can handle high cardinality, high dimensionality data at scale ○ No data/index bloat, complexity and fragility
Massive multi-tenancy	○ All customers share a common compute resource pool and all compute resources are accessed/utilized by each query
Horizontal scheduling (for a technical description, see Scalyr architecture)	○ Different from horizontal scaling, horizontal scheduling is the ability of a system to divide a single query/instruction into small discrete elements, spread those elements across the shared compute resource, and parallel process the request ○ Ad-hoc queries and complex analysis handled by shared compute cluster
Summarization/time series engine to optimize for pre-computing and visualization of repetitive queries	○ Repetitive API queries, associated dashboards and alerting on events are off loaded from the ad hoc search engine by a summary engine
Support a wide variety of data producers and shippers, as well as logs directly from applications via APIs and ecosystem integrations	○ Ability to decouple the producers and consumers of data and support each as needed with a permanent data store accessible via APIs ○ Often used as the permanent data store for Kafka events

USE CASES FOR AN EVENT DATA CLOUD

There are many use cases for an event data cloud. Some of these are met with complete SaaS offers like Scalyr's log analytics service, while others can be met through custom applications built on the event data cloud by companies using the event data cloud's APIs to power front-ends customized for the use case.

Incident management and observability applications	○ Log analytics, ad hoc and repetitive queries ○ Alerts ○ Dashboards and trending
Central repository and single source of truth	○ To resolve problems within and across systems and teams ○ To train and feed AI/ML models ○ Forensics, compliance
Performance optimization and product insights	○ System optimization and performance tuning ○ Trends in usage, pending performance problems ○ Proactive insights into customer needs
Foundation for custom applications (ie: OEM solution that is API driven)	○ Used as the analytics/query service (the 'engine' under the hood with API-driven access and integration) for SaaS analytics services, usually replacing Elastic search in machine data environments ○ Custom-developed applications specific to your business
Data-in-Motion sink and sync	○ The event data store operates as a consumer of message buses such as Kafka or Kinesis. It is an ideal sink and persistent store for these systems, storing event data for weeks, months, or years for compliance, reporting and transformation. ○ New streams of events can be produced from any data in the event data cloud. Published as new topics to Kafka or Kinesis to gain new flexibility in securely sharing data amongst internal systems, departments and even integration partners.

BUSINESS RESULTS AND BENEFITS

The technical underpinnings required to deliver a modern event data cloud are important only in that they enable several important business benefits. Since everyone touts the same benefits, including first generation log management solutions, the technical underpinnings are important to understand the likelihood of delivering on the promises, especially at scale.

TAKEN TOGETHER, THE BENEFITS OF AN EVENT DATA CLOUD FOR YOUR DIGITAL SYSTEM OR DIGITAL BUSINESS INCLUDE THE FOLLOWING:

FLEXIBILITY	○ Supports a number of mission critical use cases ○ Burstable ingest and query; it's there when you need it most ○ Structured and unstructured data supported ○ Many ways to pay: fixed monthly, burn down without penalty for unanticipated usagespikes ○ Supports short and long term data retention
PERFORMANCE	○ High ingest speed ○ Low ingest latency (<5 seconds) ○ High query speed ○ Low query latency ○ Scales to petabyte scale
AFFORDABLE	○ Low cost per GB • <$5/month per average daily GB of ingest at scale ○ COGS reduction and GM improvement • Usually 2x-10x less relative to first generation and open source solutions ○ No care and feeding required • Infrastructure, shards, storage, compute, etc. • Off-the-shelf features and capabilities • Continual product improvement ○ Predictable and controllable • No over provisioning needed • No penalties for bursts • Fixed monthly, or burndown, payment options

IN SUMMARY, WHY CONSIDER AN EVENT DATA CLOUD?

Event data is the most granular and foundational view into the health and performance of a digital service, and the event data cloud makes it easy and affordable to use this data across engineering, IT, product planning, and business planning to solve critical problems, derive fresh insights, and optimize performance.

Engineering -- and by implication, engineers -- are the most critical resource in a digital business. This is where service availability is ensured and innovation thrives. The best and brightest engineers want to develop applications that are central to the mission of the company. By contrast, their talents are under-utilized in the care and feeding of foundational support systems.

The event data cloud provides an affordable and foundational business utility to support engineers as they create new services and insights to propel the organization forward. It allows them to easily tap into the data they need, while freeing them to innovate, resolve problems and optimize performance.

Why consider an event data cloud? Because Scalyr can provide the best event data analytics engine and your engineering talent can focus on the core business and the applications that make use of the event data.