CHAOSSEARCH

A DETAILED LOOK AT CHAOSSEARCH

The Cloud Data Lake Platform for Multi-Model Analytics

Data generation, particularly the tsunami of machine-generated log and event data, is outpacing Moore's Law and existing architectures. ChaosSearch was purpose-built for today's massive-scale data generation, applying its patented indexing technology to your cloud data to enable search, SQL and machine learning workloads with infinite scale, lower cost, no data movement, and faster time to insights. Learn how it works.

CONTENTS

NEED FOR A NEW APPROACH

If a business is not data driven today, it is striving to become so. Data is the lifeblood of every organization and data scientists, business analysts, and end users rely on fast access to data to successfully perform their jobs.

Data lakes promised low cost, efficient storage and on-demand access to data. However, due to the countless number of data formats and myriad analytical tools in use, Data Operations (DataOps) and Data Engineering teams spend all their time extracting, transforming, and loading (ETL) data into different analytics platforms before the data consumer can run a single query. The data lake becomes a costly data swamp with a mess of data pipelines moving data in and out and among different analytics platforms. The result is data being duplicated, data compliance mandates being violated, and users waiting weeks or months for data to be made available. Not to mention valuable engineering resources bogged down managing a complex infrastructure and pipelines rather than contributing to innovation.

The pain is felt most acutely today with the tsunami of machine generated log and event data, which is outpacing Moore's Law and existing computer science and architectures. Consider SaaS companies, born in the cloud, that must monitor hundreds of gigabytes to multiple petabytes of system, network, and application log data every day. If data is "only" stockpiled, and at worse thrown away, valuable insights are lost—both from an operational and business perspective. As a result, organizations cannot keep environments up and running, mitigate persistent security threats, or improve customer experience.

Conventional data analytics solutions are based on technology that, at its core, is decades old. They were designed before widespread cloud adoption and when data volumes were smaller, data sources were simpler, and the number of active users requiring analytical access were few. However, to meet the demands and opportunities of today and tomorrow, storage and analytics will need to fundamentally change.

Data is Exploding From Everywhere

It used to be that data was primarily created from internal sources (i.e. CRM and Database systems), in a structured format, and at predictable volumes. Today, with the advent of mobile, cloud, and microservices—data is coming from a vast and rapidly changing array of sources, including operating systems, cloud services, applications, and mobile devices. It arrives in structured, unstructured, and semi—structured formats such as CSV, LOG and JSON, at highly variable and growing rates and volumes.

Traditional Approaches Are Broken

Due to this explosion in both data size and complexity, today's data analytics solutions struggle to maintain a balance between scale and cost. One popular solution is the open source ELK Stack (Elasticsearch, Logstash, and Kibana). It's easy to get started, but quickly becomes too complex, too labor intensive, and too costly as data volumes and sources grow. Managing such environment expansion, with all the operating systems, applications, and security provisioning, configuration and upgrades—quickly becomes a maintenance nightmare. And do not forget that scale requires load balancing, high availability, and redundancy; each aspect necessary for an enterprise-level production analytics solution.

And at the same time, "big data" offerings such as Hadoop failed to provide a legitimate alternative. The idea that analytics would be as simple as streaming raw data into a Hadoop data lake not only failed, but coined the term, data swamp. Such solutions were not designed for efficient, fast, and cost-effective data analytic access. They require difficult-to-find skill sets where each use-case and workload typically has a homegrown, "schema on read" design pattern. And if Hadoop was not the primary analytic platform, but a lake repository, it was often used to run ETL workloads into purpose-built, siloed warehousing solutions.

The Move to Cloud Services

Cloud computing is also exploding. Cloud platforms and applications are proliferating across enterprises and startups alike, serving as the IT infrastructure driving new digital businesses. The cloud wasn't even conceived of when traditional data analysis solutions were designed. Consequently, they weren't built to take advantage of the unlimited scale and capabilities of new services, like cloud object storage.

What is Needed Moving Forward

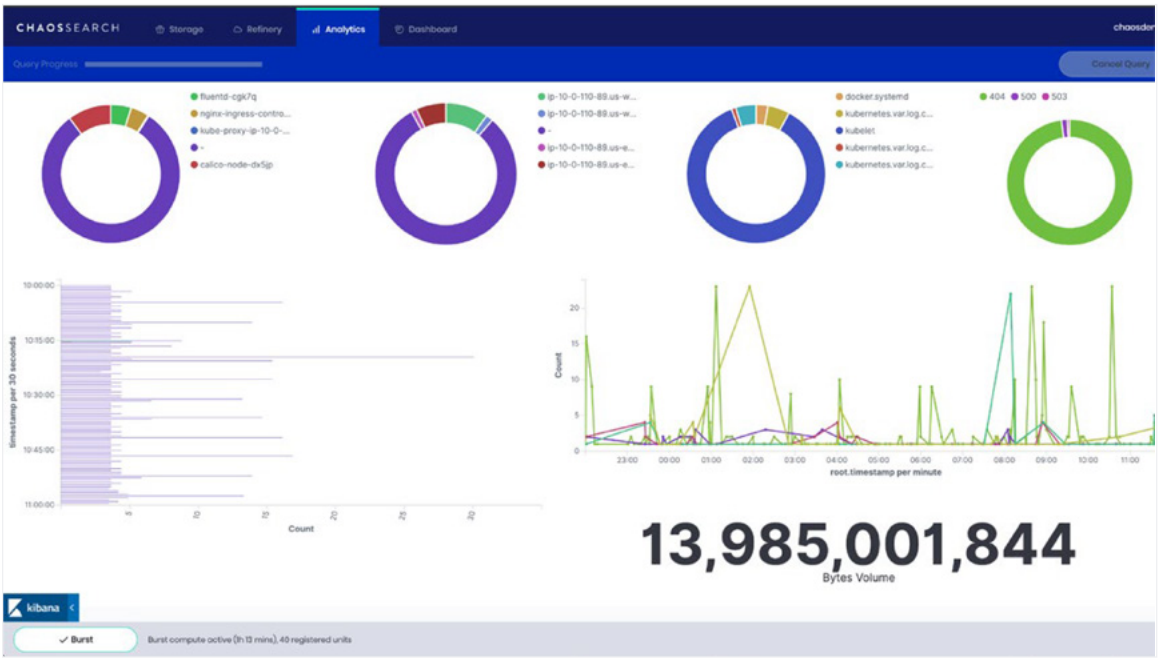
Access to data for analysis should not break the bank. And as indicated, it's now the lifeblood of business, fueling the new information age. However, with this need to grow based on more and more data, most organizations need dynamic, scalable, and efficient data analytic infrastructure, where new workloads can simply be added with one click, and payable as a subscription cloud service—as opposed to an expensive, fixed, capital investment.

> “With ChaosSearch, we are able to use a singular solution for our various logs without the hassle of managing the logging tools as well.”

Joel Snook
Director of DevOps Engineering at Blackboard

ACTIVATING CLOUD OBJECT STORAGE AS A HOT, ANALYTICAL DATA PLATFORM

Current limitations and challenges cannot be fixed by simply tweaking legacy architectures with new capabilities or simply re-written to support the cloud. These architectures are fundamentally inadequate to cost effectively address the diverse and growing needs of today's software-driven businesses. To address these shortcomings, a complete redesign is needed; a reimagination of a data access architecture via new computer science. The ideal solution would start with the best of cloud services—performance, security, known tooling, and APIs—but with the advantage and benefits of new thinking, producing disruptive pricing, and a flexible "pay-as-you-go" service.



ChaosSearch can easily handle terabytes of data from any data source

An Ideally Architected Solution Would Support the Following

Based entirely on cloud object storage: Cloud storage is built for scale. It is "home base" for keeping up with the mountain of data you need to store. The ideal solution would be able to run search and relational analytics directly in place within your own cloud storage—no data movement or physical transformation.

Delivered as a fully managed SaaS: Infrastructure would be completely delivered and managed as a cloud service so that users can focus on getting value out of data without running a database and all the complexities associated with it at scale.

Support all data types and sources: Automatically model and normalize all types of data regardless of type or source, and support data cleaning and transformations directly within cloud storage, without having to move data back and forth between multiple systems.

Work with popular tooling & APIs: Support search and relational analytics on a unified representation, integrated with popular interfaces like Elasticsearch and SQL/JDBC, and known visualization tooling such as Kibana, Grafana, Tableau, and Looker.

Instantly add workloads on the fly: Have an auto-scaling architecture that instantly scales up or down without service disruption, regardless of use case or data volume.

Available as a subscription service: Delivered as a pay-as-you-go data platform service. The flexibility of a subscription-based offering that allows businesses to consume and analyze data when they need it, without the constraints of forced volume/performance limits or legacy annual "data plans".

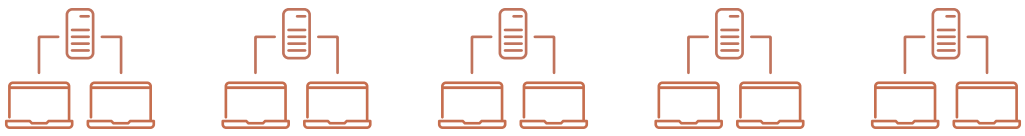
An ideally architected solution as described above eliminates all the hard work and high cost of managing a complex infrastructure, eliminates cost vs. data retention trade offs, and simplifies data access so you get faster and better insights at any scale.

LIMITS OF SEARCH AND RELATIONAL ANALYTIC SOLUTIONS

Traditional search and relational analytic solutions are based on legacy thinking and technology—designed as siloed systems, not as a cloud service; designed around block storage, not cloud object storage. And even if they have been retrofitted to run within cloud environments, where separating storage from compute is common, they are still leveraging legacy computer science algorithms and data structures.

The issues with these 25+ year old algorithms and data structures is the representation used to make databases fast. To make things fast, databases index each aspect of the raw data source. However, to make all aspects accessible and fast, databases will over-index. The more indices are utilized for speed, the more storage, network, memory, and compute are required to process them. And at some point, over-indexing actually causes performance cliffs. All of which force data and associated databases to be partitioned into separate (but connected) silos.

As just stated, these limitations are a direct result of how and where data is stored, not a result of the actual interfaces used to access it. For example, the Elasticsearch API, with ELK Kibana visualization, is a very simple and powerful interface. But to reiterate, these limits are associated with scaling and managing an ELK cluster. This goes double for relational systems. SQL/JDBC dialects have stood the test of time and are making a resurgence in recent years.



Traditional shared everything architectures are fixed and limited

Limitations of Elasticsearch and the ELK stack

The Elasticsearch database was innovative in many respects when it was first released in 2010. Its REST-based JSON interface, as well as its search and analytic capabilities at scale, were trailblazing. However, these capabilities have limitations; a result of the index technology used to support them. Elasticsearch was written in Java and uses Lucene (inverted-index) as its core technology. The issue lies in how Lucene, as a database index, scales. Though inverted indices uniquely support the full suite of text search capabilities, they have a significant achilles heel. To support fast search/analytics over the entire dataset, the Lucene index can become extremely large. For instance, it is common that a fully indexed source can be 2x to 5x the size of the raw data source. This results in more sharding where time, cost, and complexity increase rapidly.

Limitations of Traditional Relational Analytic Solutions

Relational databases have been around for decades and are extremely successful. They have gone through many iterations over the years—from the first transactional systems, to warehousing, to hybrid transnational analytic processing (HTAP), to today's cloud-orientations. However, each has the same limitations as Elasticsearch. In the case of relational systems, most utilize B+Tree based indexing (or some derivation thereof). And like Lucene, were designed to specifically run on block storage. Relational indexes do not suffer as much when it comes down to storage requirements when over-indexing. However, it does show worse with respect to the over index performance cliff. All of which results in traditional sharding techniques and the associated cost, time, complexity headaches.

Limitations of Presto-Based Multi-Connect Analytic Solutions

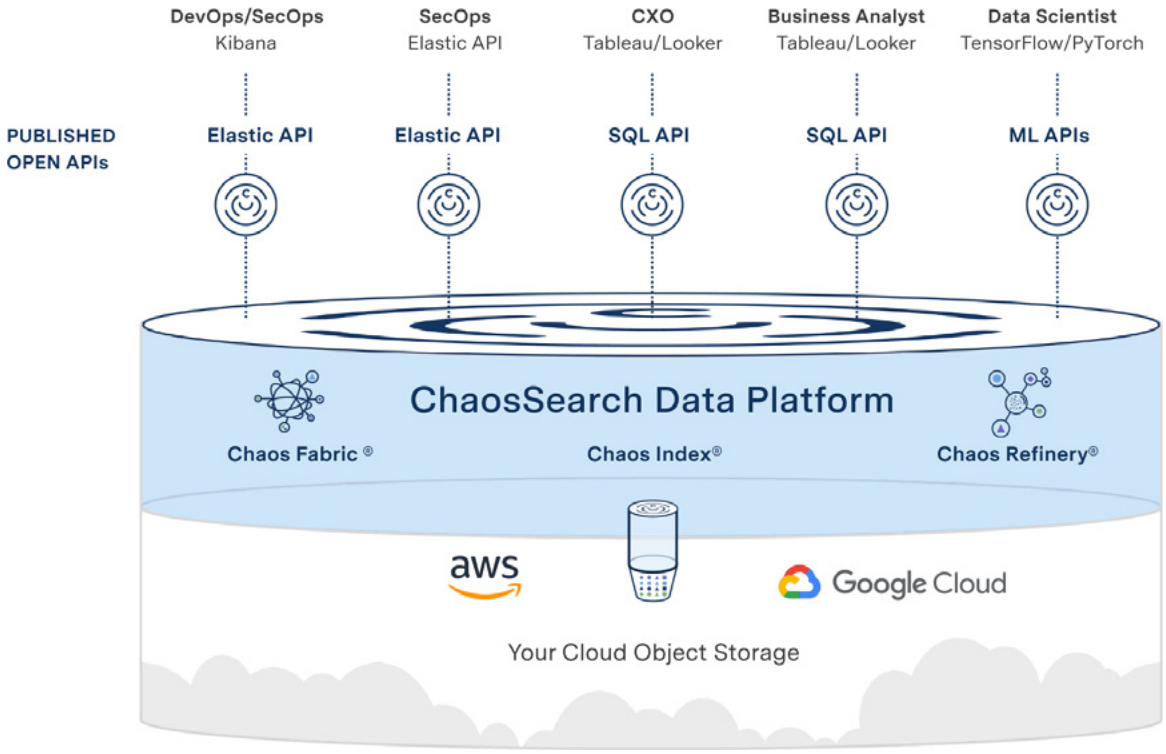
There is a new kid on the block when it comes to data analytics. Instead of moving data into a centralized solution, multi-connect database engines have recently become popular. For instance, the open source Presto engine (developed by Facebook) has answered the problem of data movement and scaling. Facebook for years used siloed MySQL sharded databases to build out their social media empire. Presto was an improvement over existing distributed query techniques, as well as the problems with their Hadoop clusters. However, multi-connect is not an actual database with transactional aspects, it is a distributed query engine (read focused); designed for discovery, not hard core warehouse analytics or, for that matter, text search (which it does not have). Presto-based solutions have all the trappings of a relational database, and in some respects, are even more complicated to set up and manage. It is common to create tons of materialized views to increase performance since a query is as slow as the slowest endpoint.

> "ChaosSearch powers our enterprise log analytics and is a critical piece of the infrastructure for processing tens of terabytes per day of our customers' log data. We've been very pleased with the performance, reliability and cost of ChaosSearch."

Josh Bosquez
Chief Technology Officer at Armor

NEW DATA LAKE PLATFORM BUILT FOR THE CLOUD

At ChaosSearch, as we considered the limitations of existing systems, we envisioned that cloud storage would be the key foundation for a simple, scalable, and cost efficient analytic data lake engine to offer search and relational analytic capabilities. Cloud object storage like Amazon S3 and Google Cloud Storage (GCS) offers near-infinite scalability, is secure, global, and designed for (11 9's) durability. Cloud object storage stores data for millions of applications for companies around the world. It is an ideal foundational platform for building a next-generation big data analytics solution.



Platform Overview

The ChaosSearch Data Lake Platform connects to and indexes data within a customer's cloud storage environment, enabling search, SQL and machine learning workloads with infinite scale, lower cost, no data movement, and faster time to insights. ChaosSearch activates your cloud object storage to become a hot, analytical data lake.

The design is based on new index technologies and an intelligent data fabric, leveraging two core cloud components:

Storage
Pure cloud object storage is where all data persists, managed and accessed. There is zero local storage used such as HDD or SSD.

Compute
Quorum of independent cloud compute resources that dynamically executes tasks like data discovery, indexing, searches, and queries.

The platform and associated services are designed to deliver simple, scalable, and inexpensive analytic access to cloud storage as well as information about it. This includes what, when, and how much data was stored. All of which enables you to search, query, and visualize data. The ChaosSearch difference is simple: Eliminate all the moving parts and challenges inherent in traditional text search and relational systems and make management easier. The mission is to streamline and automate the process and dramatically reduce the cost to store, index, search, and/or query data.

To achieve this, we first selected cloud storage platforms that are simple, scalable, and extremely cost effective: Amazon Web Service (AWS) S3 and Google Cloud Storage (GCS) (with plans to extend to Microsoft Azure Blob Storage). Next, we reinvented indexing and delivered a powerful alternative to traditional inverted (Lucene) and Columnar (Parquet) approaches. We created Chaos Index® (and its associated Chaos Fabric®)—a new distributed database that discovers, normalizes, and indexes data without human intervention. Chaos Index uniquely supports both text search and relational queries. It is optimized for cloud storage accessed over a serverless (separation of storage from compute) computational fabric. The initial focus and use case of the platform is the analysis of live and historical log and event data.

A fully managed service, the platform allows organizations to easily store, search, query, and visualize gigabytes to petabytes of data within their own cloud object storage. The platform automates the cataloging and indexing of data. With both Elasticsearch and SQL/JDBC APIs, ChaosSearch facilitates data hunting and trend analysis over large datasets by uniquely combining both text and relational analytics with popular visualization tooling.

New Technology

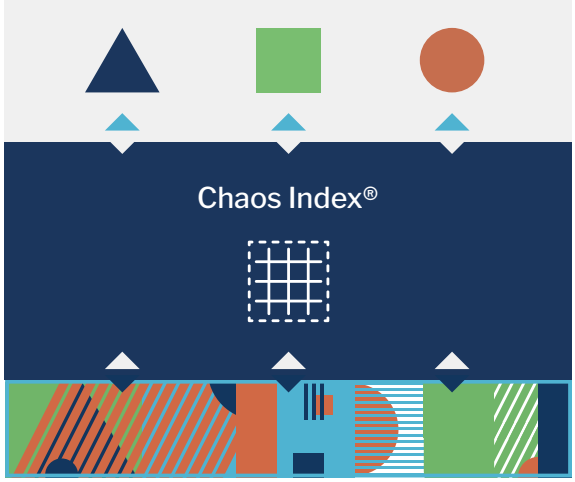
ChaosSearch consists of three central components: Chaos Index, Chaos Fabric, and Chaos Refinery. Chaos Index and Chaos Fabric work hand-in-hand, where Chaos Index enables new thinking and possibilities within the distributed computational Chaos Fabric. The Chaos Refinery allows users to virtually clean, prepare, and transform data directly within ChaosSearch with no ETL or data movement.

Chaos Index®

Chaos Index was conceived to address several major challenges associated with large scale data analytics. Any one of these challenges could be solved in isolation. However, the ability to address the issues simultaneously by leveraging one core technology was the driving principle.

Specifically, Chaos Index delivers:

- Index “all” data and “any” type of source
- Up to 95% compression at high performance
- Built-in auto normalization and transformation
- Built-in metadata and stats for optimal query planning
- Parallelizable execution model based on share nothing design
- Natively supports text search, relational queries, and machine learning
- ChaosSearch JSON Flex capability allows organizations to store all their JSON content and analyze it as if structured at different permutations and nested levels. Customers can use the Chaos Refinery to expand and explore all JSON data virtually and instantly on the fly, regardless of size or query structure.



Data indexed by ChaosSeach is compressed up to 95%

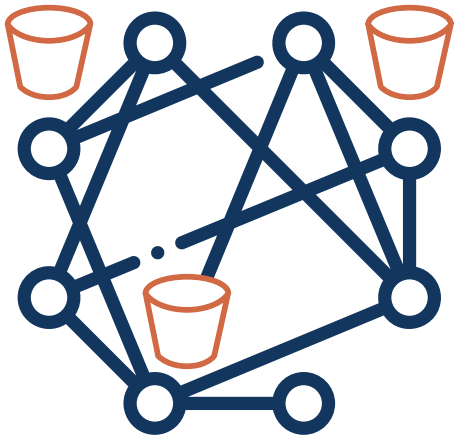
A key component of the ChaosSearch platform is the ability for Chaos Index to reduce the size of information while still fully indexing it. The ability to compress raw data sources beyond what has previously been achieved means there is significantly less storage, network, and compute resources required to back it. However, size was not the only metric designed into the Chaos Index. Speed to compress and decompress was a major aspect as well. In the case of Chaos Index, it provides compression ratios upward or greater than Gzip, with the speed of Google’s Snappy compression algorithm.

Chaos Fabric®

Chaos Fabric is based on a containerized Scala/Akka distributed architecture, oriented around Chaos Index capabilities. The Chaos Fabric is a resilient and elastic distributed actor model for orchestrating Chaos Index indexing, search, and queries. Key components include:

Key components include:

- Cost analysis metrics for distributed scheduling
- Elastic execution based on workload scheduling
- Built-in intelligent query planning and optimization
- Built-in back pressure throttling from API to/from Amazon S3
- Module front end to support multiple interfaces/APIs



Cloud storage, compute and service layers are separate but integrated.

ChaosSearch intelligently brings together the storage, compute and services layers, delivering the resources needed exactly when they are needed. And since the system is based on cloud storage, users don’t need to worry about data replication, scaling, and availability. ChaosSearch automatically indexes and stores data directly within your own Amazon S3, structured in Chaos Index’s universal format that supports both text search and relational analytics.

Chaos Refinery®

The Chaos Refinery is an in-app tool that transforms data virtually and instantly. Users publish how they want to consume and interact with data leveraging the Chaos Refinery wizard. Users can clean, prepare, and transform data directly within ChaosSearch. They can programmatically and dynamically change schema and associated data on the fly, as well as visually interact with information as needed, without the cost or complexity of additional services. The Chaos Refinery allows users to dynamically join multiple indexed data sources into one logical view to be accessed as index-patterns via Elasticsearch API, and/or tables via SQL/JDBC.

ACTIVATE THE DATA LAKE FOR ANALYTICS

Search and Analytics Directly in Cloud Storage

The simplicity of streamlining data into a lake is a powerful concept. ChaosSearch transforms data lakes into an intelligent analytic data platform. ChaosSearch built a service that allows you to store, search, and query data directly within your own cloud storage. That is a sharp contrast from current solutions, which are typically optimized for a single type of interface, forcing you to create silos for different data or use cases.

ChaosSearch took a novel, different approach, designing a data lake analytics engine that can store and process diverse types of data in a single solution without compromising flexibility or performance. ChaosSearch's approach automates each aspect of the data journey, allowing workloads to be added within minutes, not weeks or months.

Elastic, Auto-Scaling Support for All Use Cases

Conventional solutions are difficult and expensive to scale quickly, where larger environments require more maintenance. This forces businesses to invest up front in either capacity or move to a managed service with pre-bought annual data plans.

ChaosSearch is a different approach with a different architecture. Its ability to dynamically scale is a core tenet. ChaosSearch can expand and contract compute resources at any time, based on workloads. With one click, it is easy to add multi-terabyte use-cases without service interruption, and without worrying about upfront capital costs or availability of data engineers and database administrators.

New Workloads: New workloads can be increased or decreased at any time. Unlike traditional architectures where the ratio between storage and compute is a fixed formula, with ChaosSearch the relationship between storage and compute is dynamic. And since it's backed by cloud storage, this makes it possible to store indexed data at the lowest cost possible and with unlimited retention.

No Data Movement: ChaosSearch interacts directly with cloud object storage. There is no local storage leveraged. The ChaosSearch index technology and architecture allows for fast performance to cloud object storage. All transformation of data is dynamic and late materialized.

Dynamic Compute: Compute resources used for indexing data and query processing scale up or down on-the-fly as data volumes ebb and flow. And since cloud storage and compute resources are decoupled, adding or deleting compute resources does not require you to repartition data.

Query Performance: Traditional solutions have limits to scaling users since queries are in contention for the same resources. With ChaosSearch, adding new users is as easy as adding new workloads. Whether there are more requirements on indexing or queries, the ChaosSearch service dynamically meets the demand.

FASTER TIME TO INSIGHTS AND PERFORMANCE AT SCALE

Benefits of a Fully-Managed Service

All conventional databases require significant attention. It doesn't matter if it's Elasticsearch or a relational database, they require highly skilled personnel. This includes building and maintaining infrastructure, cleaning and prepping data, creating and maintaining indices, patching, cleaning up files, and on and on. This was manageable when workloads were smaller, but doesn't scale with today's growing data.

At ChaosSearch, we have built an intelligent and scalable service where users can focus on searching and analyzing data rather than tuning and managing infrastructure.

The ChaosSearch service provides the following:

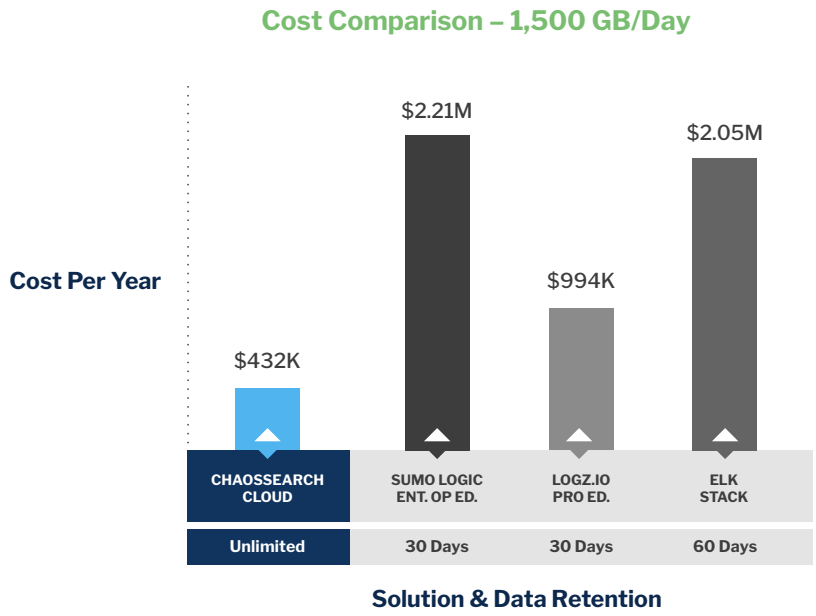
- No infrastructure (hardware or software) to manage. With ChaosSearch you don't need to think about adding, deploying or managing hardware. There's no system to manage or software to update.
- There's no tuning required. The system is intelligent and learns from its use. It conforms, adjusts, and adapts based on system workload, and number of users.
- No need to worry about capacity planning. ChaosSearch allows you to add and subtract workloads on the fly. You are not forced into an upfront annual data plan with expensive upfront costs.
- Want to know the cost of your query before you run it? ChaosSearch's dynamic query optimization allows you to determine query cost up front based on the size of the query and response time required.
- Scaling comes easy. ChaosSearch allows you to burst, as well as enable auto-scaling to match compute capacity to query load, without manual intervention.

> "Our SRE teams used to struggle with managing the vast amount of logs it takes to support millions of users in real time in a consistent manner across all our product lines. With ChaosSearch, we are able to use a singular solution for our various logs without the hassle of managing the logging tools as well."

Joel Snook
Director, DevOps Engineering, Blackboard

Architected for Cloud Consumption

There is no need for expensive, upfront data plans. Yesterday's model of licensing and/or purchasing software and hardware as an upfront CAPEX doesn't make any sense in the era of the cloud. ChaosSearch is delivered as a fully-managed service. Due to a unique architecture and approach, ChaosSearch is up to 80% less expensive than comparable search or relational analytics solutions.



ChaosSearch is up to 80% less expensive than comparable solutions.

WHAT'S NEXT

The ChaosSearch service is backed by powerful new technology and architecture designed for multiple use cases and deployment options. Initial focus is on the inherent challenges of live and historical log and event analysis where existing solutions are truly cost prohibited. However, this platform has uniquely married storage and analytics with a data lake philosophy, where analysis can be performed in-place, without the heavy lift, and at an extremely low cost. The possibilities are endless!

> "We spent too much time & money maintaining our ELK stack and started to look at managed services to handle our logging events. We could not justify the costs we were getting back from other solutions until we found ChaosSearch."

Jason Standiford
VP Engineering at Revinate

ABOUT CHAOSSEARCH

ChaosSearch empowers data-driven businesses like Blackboard, Equifax, and Klarna to Know Better™, activating the data lake for analytics. The ChaosSearch Data Lake Platform connects to and indexes data within a customer's cloud storage environment, enabling search, SQL and machine learning workloads with infinite scale, lower cost, no data movement, and faster time to insights.

Whereas all other solutions require complex data pipelines consisting of parsing or schema changes, ChaosSearch indexes all data as-is, without transformation, while optimizing for both data size and performance. Through open APIs, users continue working in the tools they know and trust, such as Kibana, Elastic, Looker, and Tableau. ChaosSearch overcomes the cost and complexity of competitive solutions, unlocking data retention limits and reducing costs up to 80%, delivered in a fully managed service.

The Boston-based company raised $40M Series B in December 2020 and is hiring to support its hyper growth.

For more information, visit ChaosSearch.io or follow us on Twitter @ChaosSearch and LinkedIn.

info@chaossearch.com | www.chaossearch.io