Enterprise Strategy Group | Getting to the bigger truth.™

ESG WHITE PAPER

A Data Platform that Delivers Insights at Scale Across Your Cloud Data

Redefining time to value at a fraction of the cost

By Mike Leone, ESG Senior Analyst; and Josh Clark, ESG Research Associate

July 2020

This ESG White Paper was commissioned by ChaosSearch
and is distributed under license from ESG.

Contents

Introduction

Organizations continue to struggle with maintaining, managing, and utilizing their data to transform their businesses. The distributed nature of data, the increased number of places where it's generated, the different structures and change rates, and the never-ending growth is forcing the hands of businesses to look for help. As data sets get larger and more complex, the cost to work with them follows suit, creating a cost versus scale tradeoff that is untenable. Data-centric enterprises are all too familiar with the maintenance nightmare that comes with managing these types of environments. And while some use cases, such as log analytics, have a relatively low barrier to get started, it quickly becomes time consuming, complex, and highly labor-intensive as more organizations prioritize integrating data across the business from different systems to deliver a more comprehensive view of the business with actionable insight.

The story arc of data lakes began as a solution for these issues, but quickly failed to provide organizations with the right capabilities. What started as a data lake that appealed to DIYers, quickly turned into data swamps due to inabilities to store, access, and process large-scale and diverse data sets, never mind the growing challenges associated with management, maintenance, the evolving list of tool integrations, and the growing number of users eager to get their hands on data. It became a cost burden to find the right skill set needed to manage and maintain an optimal workflow for data analytics on a data lake. Enterprises quickly realized that solutions like Hadoop were not as cost effective, time effective, or efficient in returning value based on the investment as they had once thought.

Taking Back Control of a Data Lake

Organizations are recognizing the opportunity to transform their data swamps from value inhibitors into what they thought they were buying into the first time around: a value-enabling data lake. Organizations are seeking a solution that embraces the diversity and distributed nature of data with scalable cloud services that enable preferred tool integrations without breaking the bank. ESG research shows that a modern business requires:

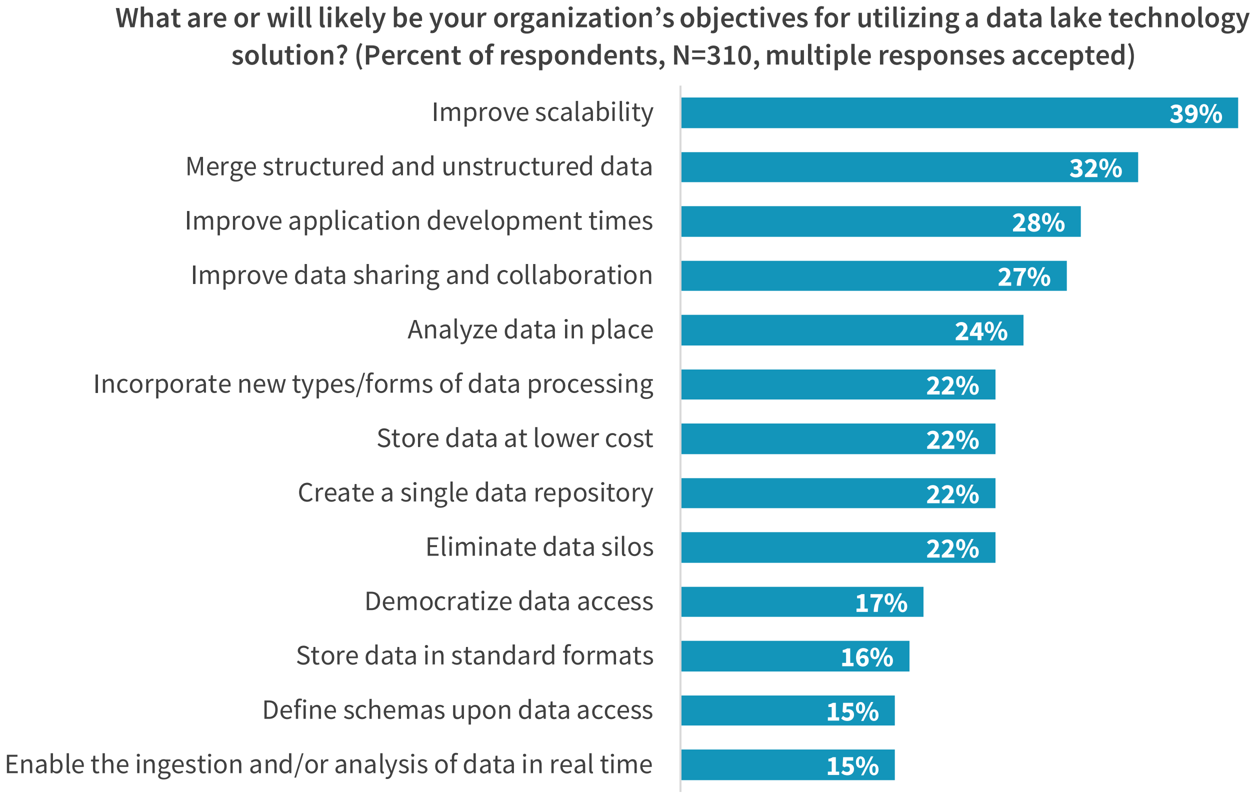
- On-demand scalability. As shown in Figure 1, the most cited objective for organizations currently using or looking to use a data lake is improving scalability.[1] An ideal solution has automatic scaling that ensures the right amount of resources is available to support any use case, data volume, or simultaneous end-user count.
- Cloud. When identifying the top areas of data analytics investment over the next year, nearly 2 in 5 organizations cited cloud-based analytics.[2] With many organizations piling data into their cloud-based storage platforms, running advanced analytics in place without the need to move data across different platforms is preferred.
- Managed service. More than half (53%) of organizations currently utilize or plan to utilize an analytics managed service. In other words, organizations want to focus on gaining value from data, as opposed to constantly tinkering with compute and storage based on dynamic workload demands.
- Data diversity. 60% of organizations utilize a mixture of unstructured and structured data. As organizations are more commonly trying to leverage all types of data, supporting all data types and sources is essential. The solution in use should automatically normalize all data types and sources directly in cloud storage without moving the data across multiple systems.
- Third-party tool integration. As organizations look to modern approaches to data lakes, embracing preferred tools and technologies will best enable businesses to quickly ramp up their data-centric initiatives. With 32% of organizations leveraging real-time/streaming analytics on their data lakes and 27% integrating dashboards/

[1] Source: ESG Master Survey Results, *The State of Data Analytics*, August 2019. All ESG research references and charts in this white paper have been taken from this master survey results set unless otherwise noted.
[2] Source: Master Survey Results, *2020 Technology Spending Intentions Survey*, January 2020.

visualizations, tool and API integration support will give data teams peace of mind in selecting the ideal data lake engine.

- Pay-as-you-go pricing. 31% of organizations state cost is a challenge when deploying or supporting a data lake. Existing data lakes are too expensive, using fixed annual costs centered on performance delivery. Organizations are looking for a subscription-based model that lowers the barrier to adoption, simplifies onramp, and promotes data access and analysis.

Figure 1. Top Objectives for Utilizing a Data Lake



Source: Enterprise Strategy Group

What's Available Today and Where Does It Fall Short?

As vendors take different approaches to helping organizations refine and evolve their data lake objectives with modern architectures, the market as a whole still has several limitations. Most solutions lack at least one major capability. Some solutions that worked well in the beginning are struggling to handle the speed and scale demands on the system. Traditional relational analytics solutions that introduced the idea of hybrid transaction/analytical processing (HTAP) in the cloud had been deemed a success, but many businesses are now questioning the ramp-up time, with implementation and customization receiving warranted scrutiny. Additionally, skills gaps and challenges with ongoing performance optimization and efficiency with increased concurrent user counts are proving to be a losing battle.

Another aspect of shortcomings in many solutions is the handling of diverse workloads. Some solutions are one-dimensional in that they are designed to handle purely historic data, real-time data, a data warehouse workload, or ad-hoc querying. For example, several point solutions in the search/log analysis space don't enable an effective way to pass data access over to a data warehouse solution to incorporate business intelligence. This creates unnecessary confusion and leads to poorly managed and structured data repositories. Rather than unique siloed solutions for each type of analytic need, a single system leveraged by a variety of end-users across many different workloads is needed.

Delivering on the Data Lake Promise with ChaosSearch

ChaosSearch helps businesses move past data lake limitations by providing an effective way to utilize data for analytics at scale. With years of experience transforming enterprises into data-centric businesses, ChaosSearch's mission is to deliver on the promise of data lakes. The company offers a managed service that turns cloud object storage into a scalable, searchable analytics engine, streamlining and automating the storing, indexing, searching, and querying of data. And with a growing list of APIs, businesses gain an effective way to find more data and conduct advanced trending with detailed visualizations.

ChaosSearch introduces a new and differentiable way for organizations to interact with data using a new approach to indexing along with an intelligent data fabric, both of which utilize cloud object storage and compute. Cloud-based object storage handles the persistence, management, and access, while compute satisfies quorum necessities to handle discovery, indexing, searching, and queries.

- Chaos Index – ChaosSearch rethought how indexing should work, inventing a new approach that utilizes a new distributed database to discover, normalize, and index data autonomously. Chaos Index provides a multi-model universal data format that reduces the size of information while still fully indexing it. It supports both text search and relational queries across a unified data set consisting of mixed data types and sources. And this is all done through a small footprint with a high compression rate that increases performance, delivering a highly optimized, cost-effective platform that enables a new generation of data-centricity for businesses.

- Chaos Fabric – Based on a distributed architecture, Chaos Fabric delivers containerized orchestration of Chaos Index's functionality of indexing, searching, and querying. Distributed workload scheduling enables elasticity with cost analysis metrics. And intelligence is infused throughout Chaos Fabric to help plan and optimize querying with API throttling to ensure budgets stay in check.

- Chaos Refinery – This tool enables end-users to publish consumption and interaction data models using an intuitive wizard that instantly and virtually transforms data in a customized way. Within the ChaosSearch platform, Chaos Refinery enables end-users to programmatically clean, transform, and prepare data, whether looking to change schemas or interact with real-time data on their terms. The outcome is a single, logical view of data based on dynamically joining multiple data sources using data index patterns from existing tools (e.g., Elasticsearch/Kibana, SQL/Looker, etc.) via APIs.

For ChaosSearch, reliable speed, scale, simplicity, and cost-savings serve as pillars to its analytics approach.

Speed

ChaosSearch implements a data lake built for real-time ingest and augmentation of existing analytics workloads. Data does not need to move from its existing cloud object storage platform, enabling customers to overcome their existing hurdles when it comes to consolidating data silos. ChaosSearch has the ability to leverage a customer's existing Amazon S3, so the speed at which organizations can get started is near instant. And since enterprises do not need to purchase new local storage, there is no need to traverse multiple platforms to ensure optimal and ongoing integrations.

While quickly ramping up the platform is a speed highlight, the next speed benefit is the speed at which organizations can gain insight and value. ChaosSearch understands time is always of the essence, and therefore, delivers a managed service that removes the need to worry about ongoing maintenance, capacity planning, and resource availability. End-users can add their desired workloads on the fly and, with auto-scaling, ensure a right-sized infrastructure to support the dynamic nature of those workloads. By virtually eliminating manual infrastructure management and intervention to match compute to a workload, a new level of time to value is delivered, enabling a more efficient working environment that allows end-users to focus on insights and outcomes.

Query performance is one of the most critical aspects of speed, bringing up questions such as: How long will it take a query to execute? How will additional end-users using the platform impact query responsiveness? With no predesignated limits for how queries distribute resources, the highly compressed footprint enables queries to easily scale to address the performance demands of end-user activity. And with ChaosSearch claiming a size and speed advantage over the likes of Gzip, the possibilities are appealing to many who have been forced to make size versus performance tradeoffs. The dynamic nature of the ChaosSearch system in meeting new workload demands ensures existing query performance is not impacted as new workloads are run on the system. In other words, end-users can leverage optimized query planning and execution effortlessly and efficiently at any scale.

Scale

To address the management and maintenance challenges organizations face with constantly evolving and growing environments, ChaosSearch offers dynamic scaling. Offering auto-scaling, the platform can expand and contract compute resources based on current workloads. This completely eliminates the need to repartition data in order to forcibly decouple, add, or delete compute resources. Chaos Fabric automatically orchestrates the scale of compute resources as the volume of data increases or decreases. Compute capacity intelligently matches query load without manual intervention, learning over time to eliminate tuning.

> **"Unlike traditional architectures where the ratio between storage and compute is a fixed formula, with ChaosSearch the relationship between cloud object storage and compute is dynamic."**
>
> --ChaosSearch

Simplicity

With ESG research showing management complexity is one of the greatest challenges when deploying or supporting a data lake, ChaosSearch has eliminated the need to manage infrastructure. The highly skilled (and highly expensive) talent traditionally required to maintain an analytics infrastructure can be turned to more strategic initiatives because ChaosSearch is a fully managed service; simply layer an intelligent data analytics engine on top of an existing S3 object storage bucket. ChaosSearch built a data lake analytics engine that stores and processes different types of data in a single solution without compromising flexibility, scale, or performance. And because of the implementation simplicity and management simplicity, the barrier to add new users (and their preferred workloads) to the platform is low.

Cost

In order to avoid the typical trend of organizations overpaying for expensive infrastructure that needs to be budgeted annually, ChaosSearch handles cost differently. Specifically, there is no need to get caught up in capacity planning with annual data plans that carry heavy, upfront costs. Utilizing the technology does not force organizations into upfront annual data plans because of how dynamically the system can add or subtract workloads. ChaosSearch did not want organizations to lose out based on massive upfront investments without even having a chance to maximize usage. The approach is purely pay-as-you-go, which, in the case of analytics, becomes drastically cheaper than traditional analytics pricing models.

ChaosSearch removes large upfront costs for onsite storage, maintenance, scaling, and capacity. By seamlessly integrating with an existing cloud object storage solution, there is no need to purchase new local storage. This is critical as what typically starts out as a small, one-dimensional workload can quickly balloon into a highly complex and constantly growing data set of mixed structure and change rate. ChaosSearch has cracked the code at delivering a cost-effectice, scalable analytics platform by rethinking how indexing works. Chaos Index reduces the size of an organization's data footprint meaning less storage, less data movement, and less compute to search it. And it leads to drastically lower costs without compromising any features or functionality.

Where ChaosSearch Is Going Next

With cost and scale being prohibitive to many today, ChaosSearch has initially focused on the challenges of historical and real-time log and event analysis. And the great news for businesses just getting started with a data lake is that log and event analysis consists of highly structured data, meaning it's a use case that can serve as a launching pad to more in-depth or complex analytics use cases. ChaosSearch was founded with the idea that it is and always has been about more than log analytics. The opportunity of incorporating more data, from different data sources, of different structures, with different change rates is where ChaosSearch is headed as its team works to create a comprehensive data platform using a data lake philosophy.

With an API-centric platform, expect more connectors to the plethora of tools in the analytics and business intelligence space. While ChaosSearch supports AWS today, multicloud support for GCP and Microsoft Azure are on the way. In the future, a universal, multi-cloud layer will be used to enable single pane of glass management and access across clouds. Additionally, ChoasSearch will be offering a VPC deployment alternative in the near future.

As ChaosSearch's data lake technology becomes even more advanced, helping organizations make better business decisions faster than ever before continues to be the ultimate goal.

The Bigger Truth

The promise of a data lake quickly caught the attention of everyone. But what started as promise and potential quickly turned into disaster. Management complexity and cost were untenable to a majority of organizations. Internal champions continued to feel pressure to show value from these massive investments that struggled to meet the lofty expectations of the business. Even with minimal value being delivered from the first pass of data lakes, organizations still desired to be more data-driven. They turned to pointed tools based on the use case or data structure. This introduced a new wave of management complexity, and as data kept growing, budgets all but burst into flames. A rethinking of the data lake was required.

ChaosSearch is delivering on the initial promise of a data lake with a new approach that delivers ubiquitous data access to all data no matter the size, scale, or speed. This empowers businesses to gain insight and take action by rapidly ramping up and promoting data access and analysis to more people; forget about managing the infrastructure; focus on the data; and do it all at a fraction of the cost of traditional approaches with virtually no future limitations based on speed, scale, or cost.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.

www.esg-global.com ✉ contact@esg-global.com ☎ 508.482.0188