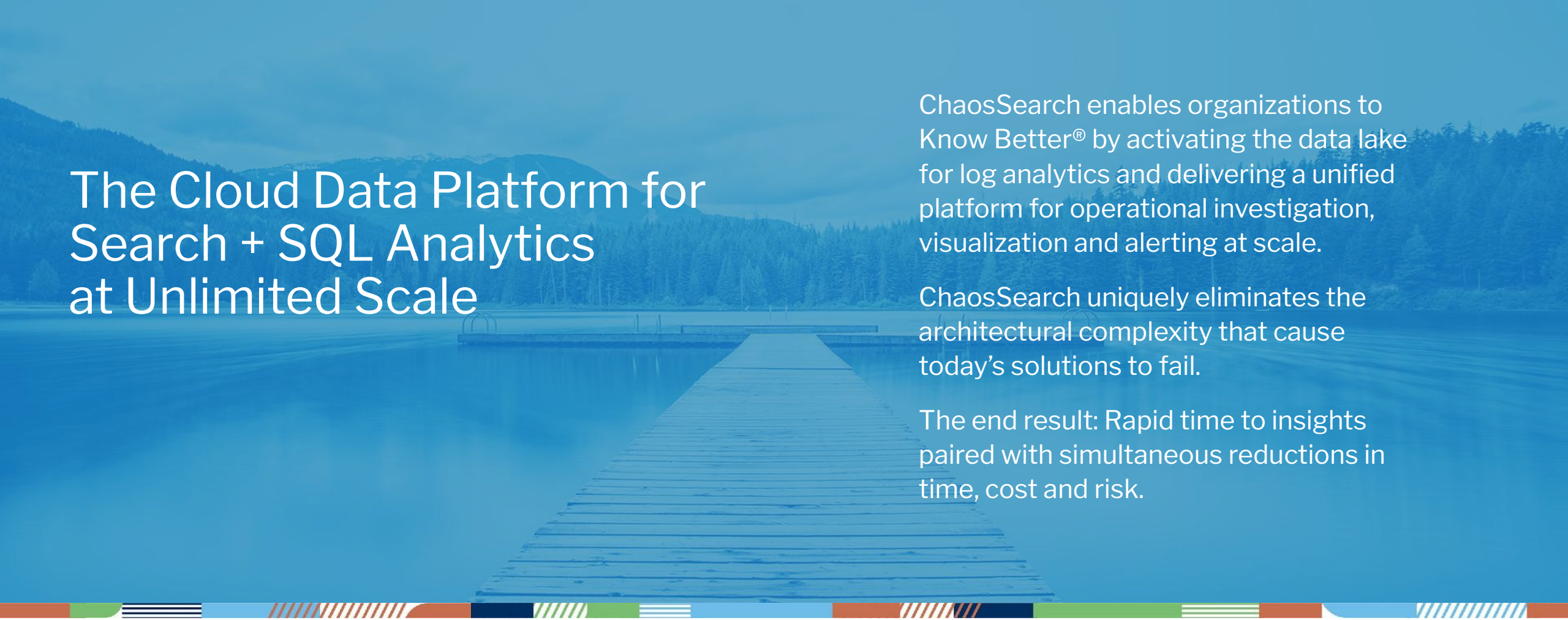
Activating Data Lakes for Analytics at Scale

CHAOSSEARCH

Greg Goldsmith
https://www.linkedin.com/in/gregory-goldsmith/

& Dave Armlin



The Cloud Data Platform for Search + SQL Analytics at Unlimited Scale

ChaosSearch enables organizations to Know Better® by activating the data lake for log analytics and delivering a unified platform for operational investigation, visualization and alerting at scale.

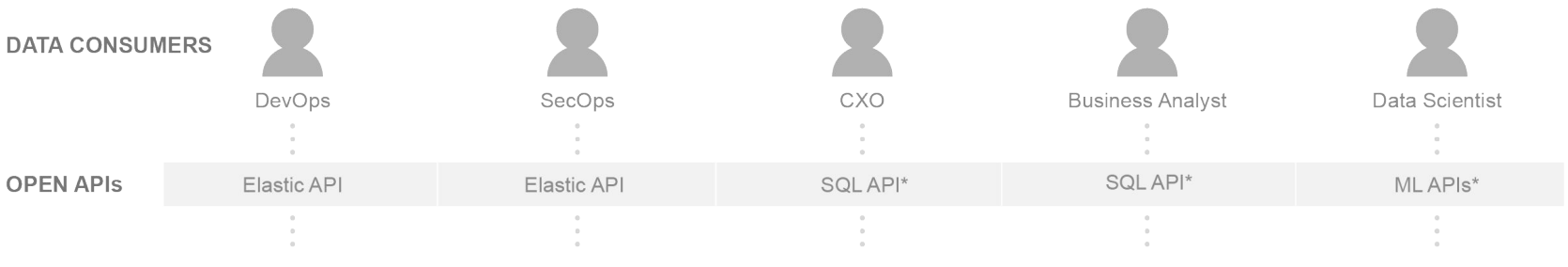
ChaosSearch uniquely eliminates the architectural complexity that cause today's solutions to fail.

The end result: Rapid time to insights paired with simultaneous reductions in time, cost and risk.

FEATURED CUSTOMERS

EQUIFAX · Blackboard · bai communications · REVINATE · Agilence · ARMOR · Klarna. · Digital River

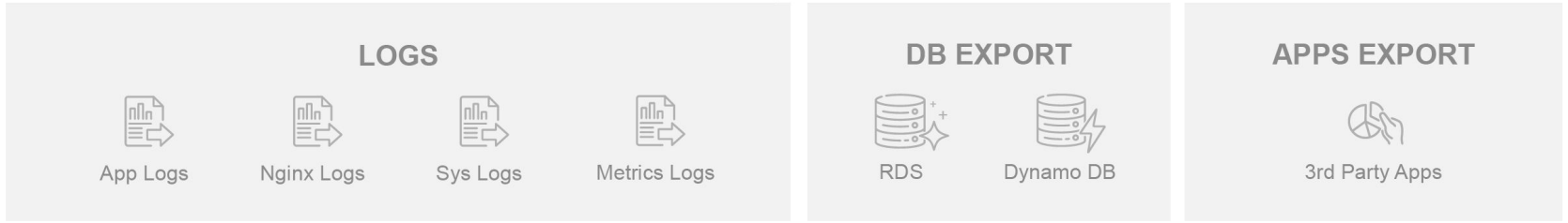
Analytics at Scale: Complex, Costly and Getting Worse



"80% of analytic workers' time is spent getting the right data, to the right place, at the right time".

- 20% searching for right data
- 37% preparing into right place and format
- 24% protecting and governing
- Using 4 to 7 different tools, adding to the complexity
- With 44% of workday spent on unsuccessful data activities

https://www.datadoghq.com/product/log-management/

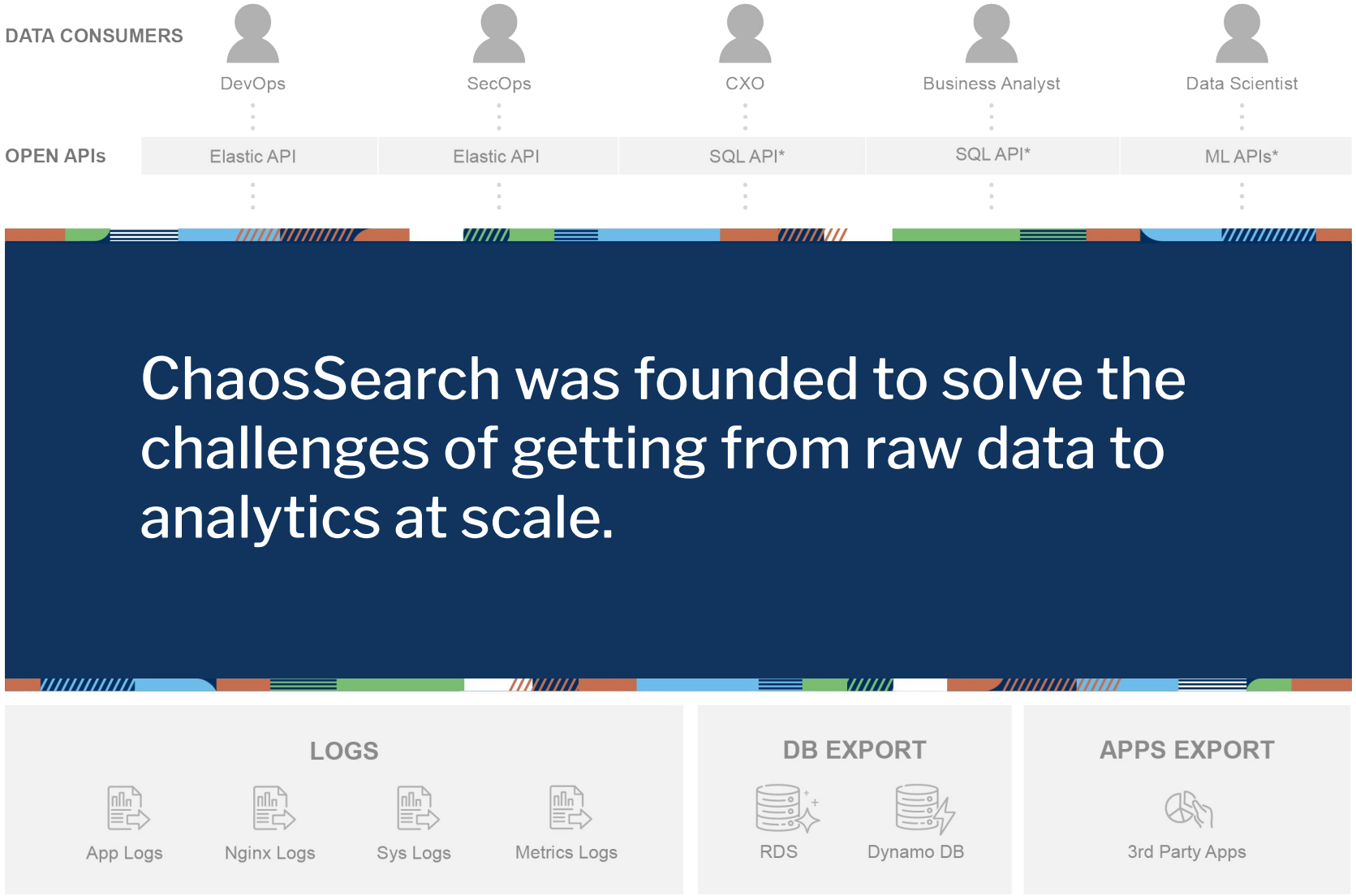


Source: IDC

MULTIPLE MODES OF ANALYTICS

INCREASING VOLUME, VELOCITY, AND VARIETY OF RAW DATA

The ChaosSearch Cloud Data Platform



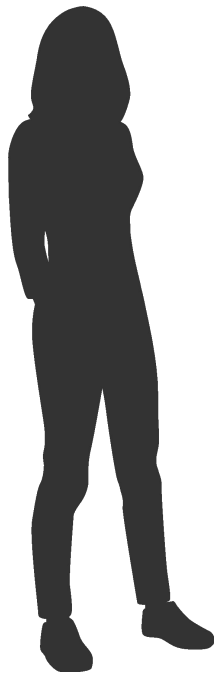
MULTIPLE MODES OF ANALYTICS

INCREASING VOLUME, VELOCITY, AND VARIETY OF RAW DATA

Operational Analytics Start Simple…

We need data to make better informed decisions!

CXO



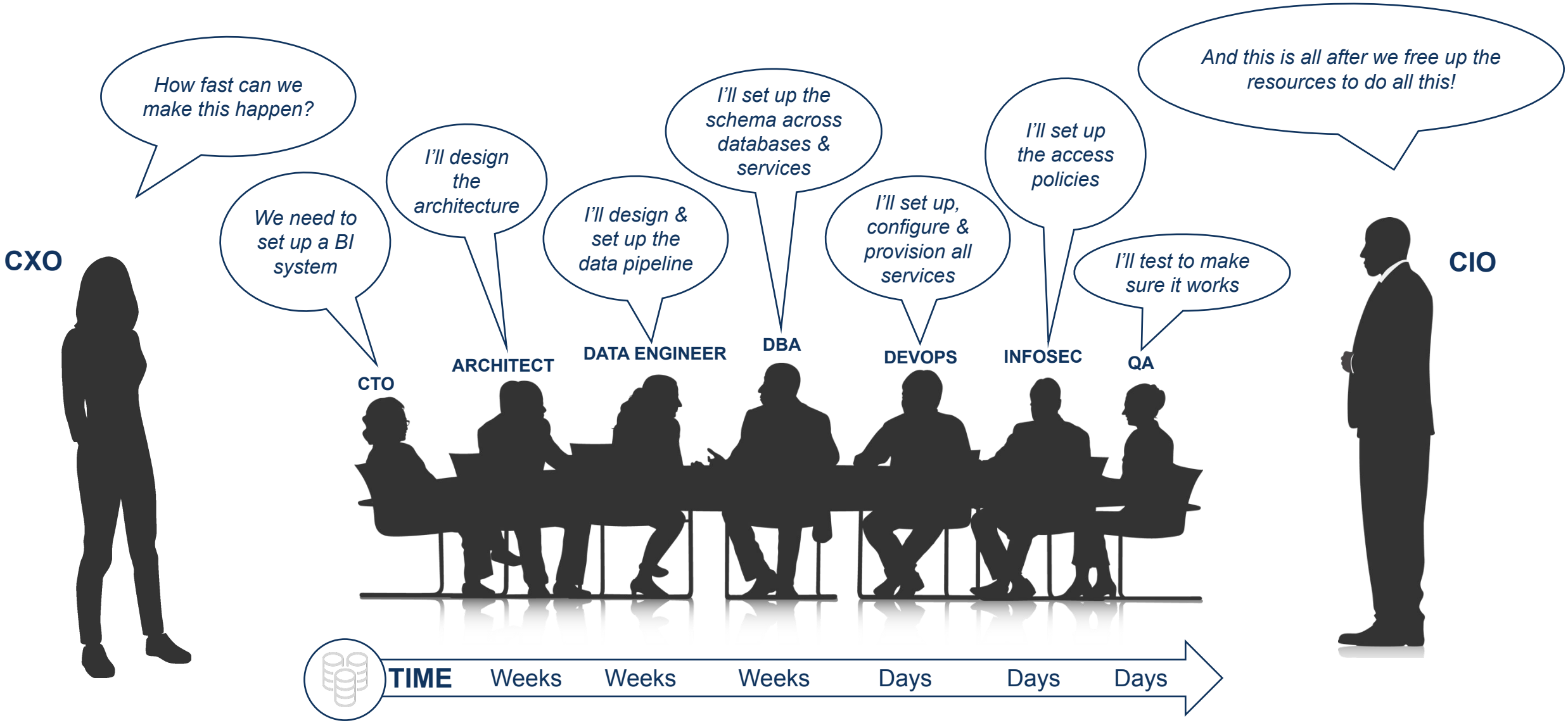
We can instrument all our systems to log everything!

DevOps

We can cheaply store them all in a cloud-based data lake!

Architect

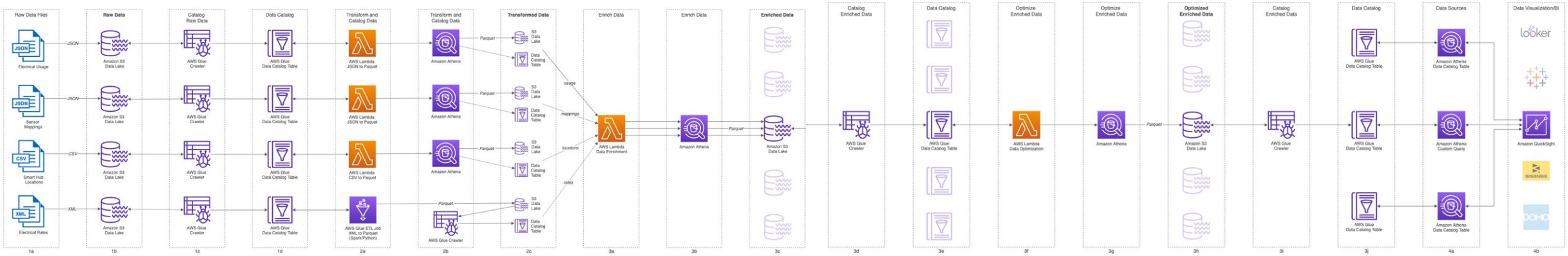
But Quickly Get Complicated



Architectural Complexity is the Root Cause

Why are time, cost and risk all increasing?

Because the complete solution looks like this…



TIME

- Effort of Planning and Implementing Each System and Process for Production Deployment

COST

- Direct Cost of Each System and Resource
- Indirect Cost of Operating and Maintaining Them

RISK

- Each is a Point of Failure & Vulnerability
- Any Change or Downtime Impacts Entire Pipeline

But Scale is the Breaking Point…
With Today's Cures Forcing a Trade Off

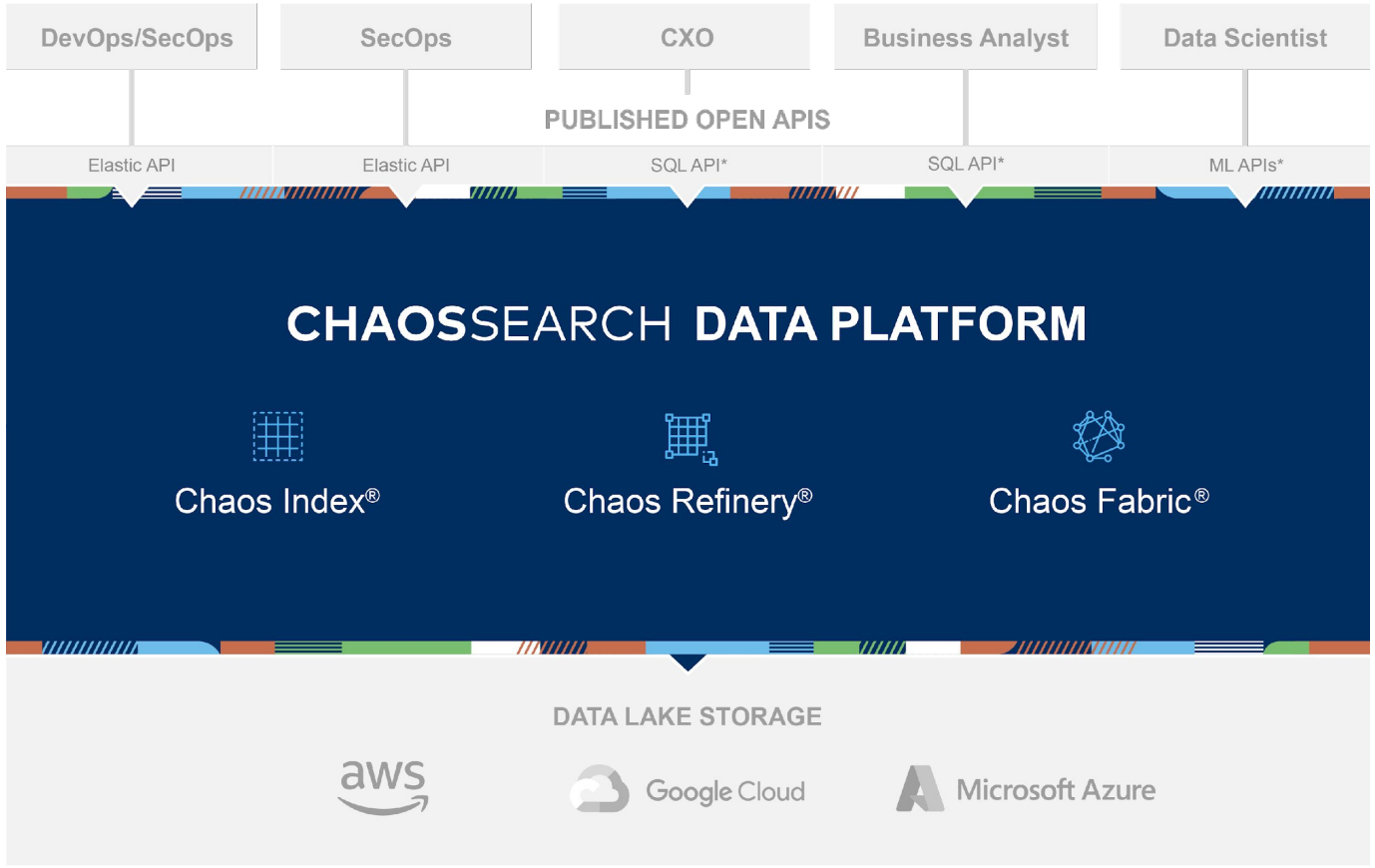
- Reduce the Amount of Data
  - Retention
- Reduce the Performance
  - Accept Slow Downs
- Reduce the Reliability
  - Accept Failure and Downtime
- Reduce the Flexibility
  - Limit What Can be Answered

VS.

Existing Solutions End in the Same Loop

- All other approaches are dependent on adding data movement into single-purpose, partitioned structures and dedicated systems
- Complex and inefficient data pipeline processes "collapse under their own weight at scale"
- Bias is introduced to data from the very beginning - inherent to the data pipeline process
- Structures are paired with complex SSD persistence and/or transient in-memory caches
- Resulting in constant tradeoffs of performance or scale

ChaosSearch Activates Your Data Lake for Search, SQL and Alerting at Unlimited Scale

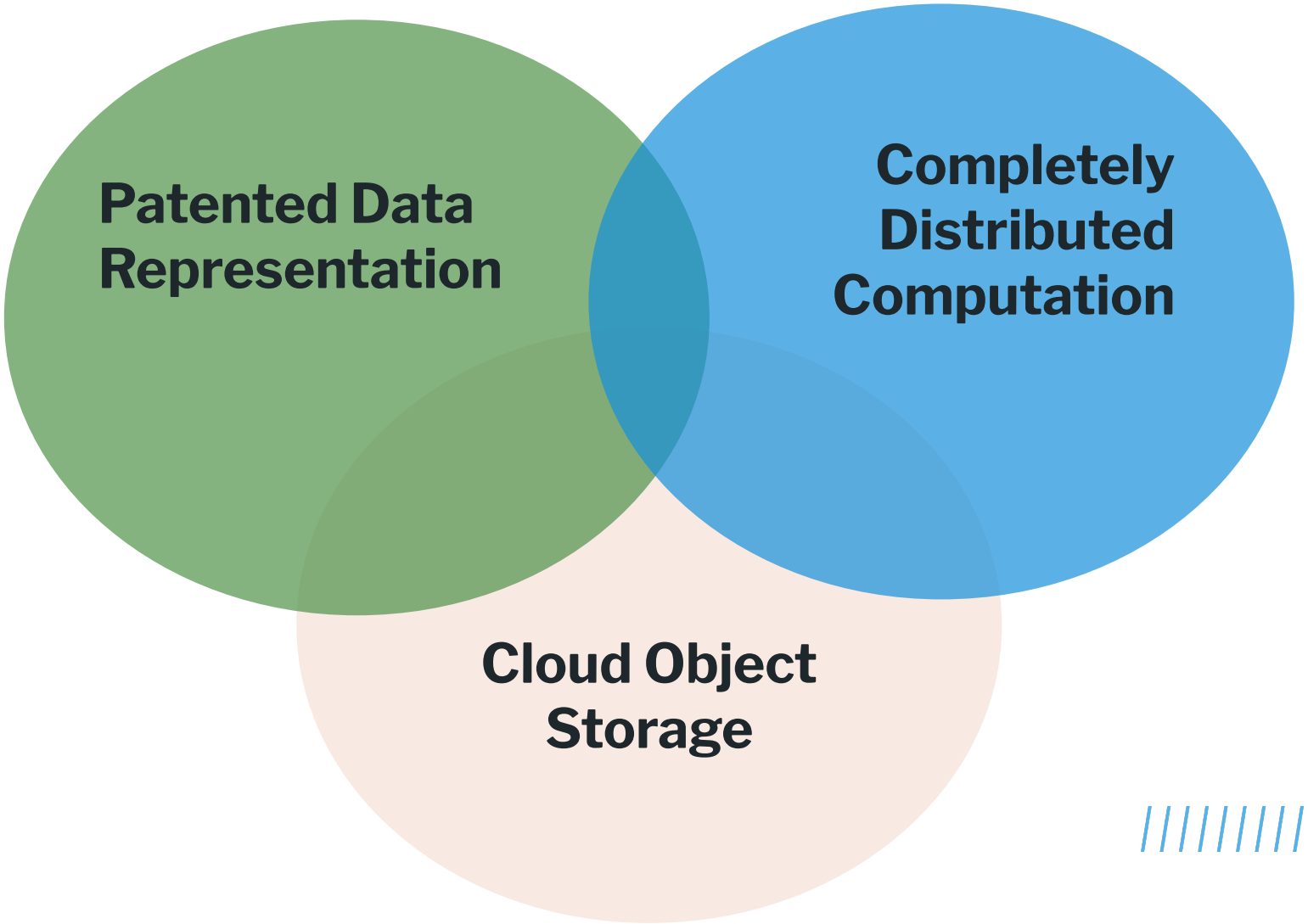


Simply ingest raw data and get instant insights out

Architected from the ground up to **permanently eliminate the layer upon layer of complexity** that is built into all other data & analytics platforms.

The resulting game changing simplicity enables unparalleled flexibility in analytics at scale while **simultaneously reducing time, cost AND risk**.

Transform Your Cloud Object Storage into a Hot, Analytical Data Platform



STORE

Ingest raw semistructured data into your existing cheap, reliable and infinitely scalable cloud object storage

INDEX

Model into a lossless, yet highly compressed, data representation
… that never leaves your storage

REFINE

Publish virtual data views for enrichment & governance
…with no data movement

ANALYZE

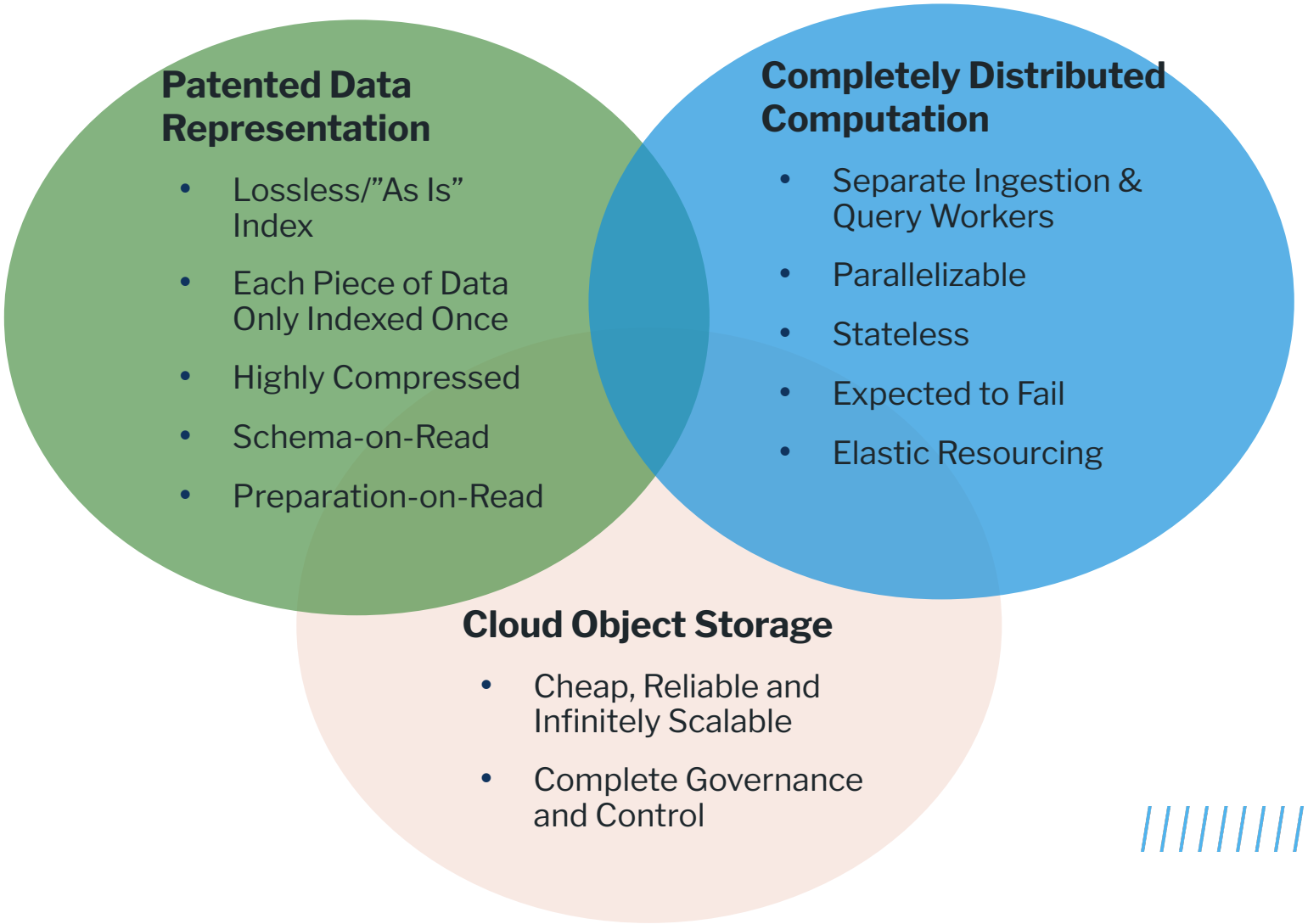
Use your tool of choice for

- Log Analytics
- Exploratory BI
- Continuous Alerting
- and Anomaly Detection*

COMPUTE

Autoscaled compute fabric for highly parallelized ingest and multi-model query at limitless volumes

Transform Your Cloud Object Storage into a Hot, Analytical Data Platform



- Every Query Goes Back to Your Cloud Object Storage
- With the Performance of Schema-on-Write SSDs
  - On first and every read
  - No pre-persisted storage or transient memory caches
- Search, SQL and Alerting in a Single Engine
  - No data movement

“With the move to ChaosSearch, 98% of all operational burdens have been lifted from us, allowing us to focus on Blackboard-specific tasks.”

Joel Snook, Director, DevOps Engineering

Blackboard®
NOW PART OF ANTHOLOGY

Optimize Operational Log Analytics

Replacing Elasticsearch or AWS OpenSearch for log analytics at scale

CloudOps/DevOps

- Unlimited retention to optimize troubleshooting and performance of increasingly complex cloud architectures
- Better log coverage to shorten time to resolution
- Eliminate administrative toil, reduce operational costs

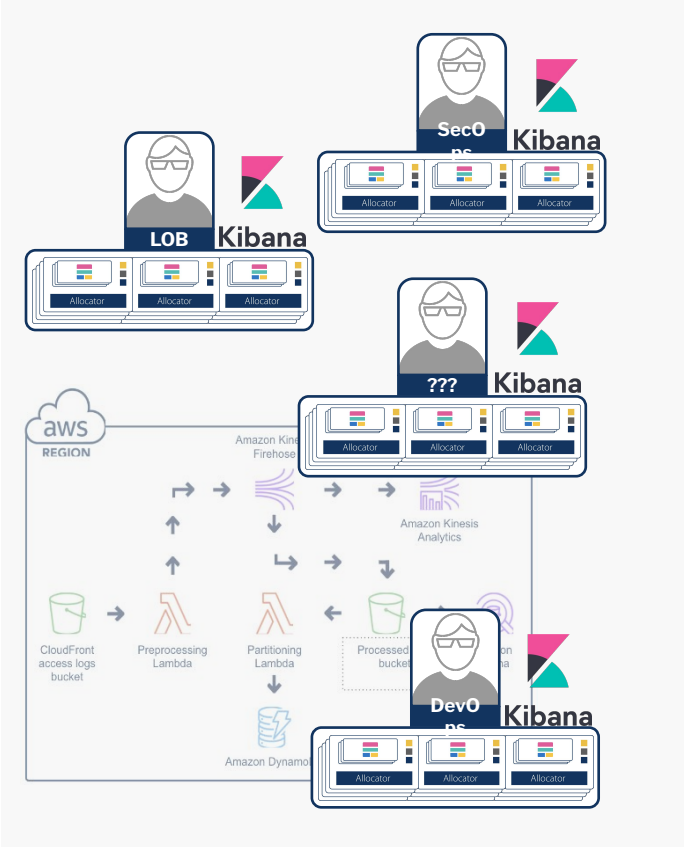
SecOps

- Affordable long-term retention for in-depth forensics
- Centralize logs in a security data lake for end-to-end visibility and monitoring
- Simpler, more cost-effective compliance

Log Analytics

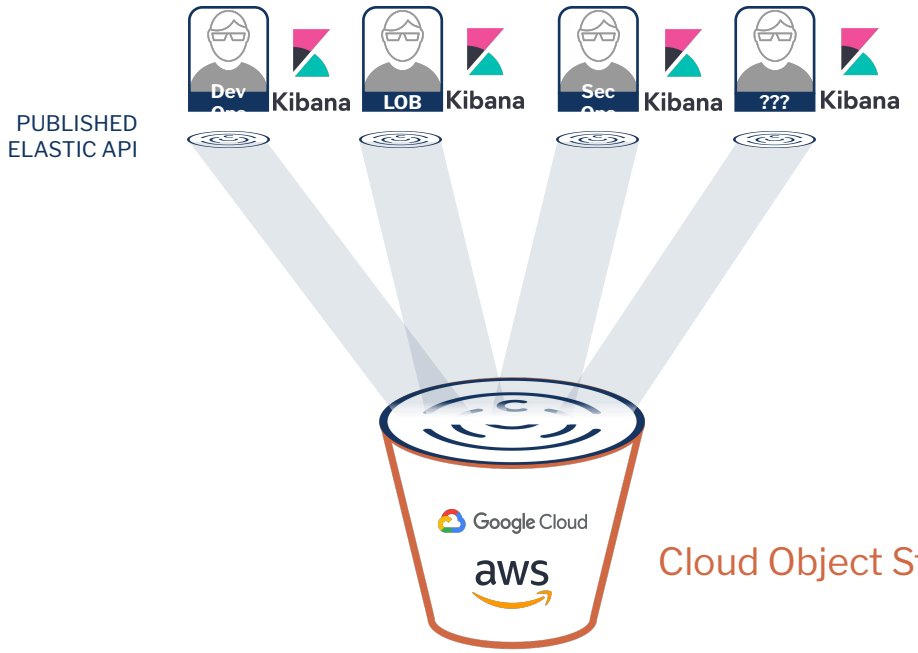
Replacing Elasticsearch or AWS OpenSearch for log analytics at scale

Before: Elasticsearch (ELK stack)



- Limited retention
- Expensive to provision for spikes
- Lags & downtime created by instability at scale
- Management and configuration challenges
- Breaks if log schema changes
- Multiple data silos created due to the limits above

With ChaosSearch



One unified data plaform

Unlimited scale and retention.
Save up to 80% on Managed Service with 99.99% uptime.

Search + SQL are Pervasive, but Siloed and Limited in Scale

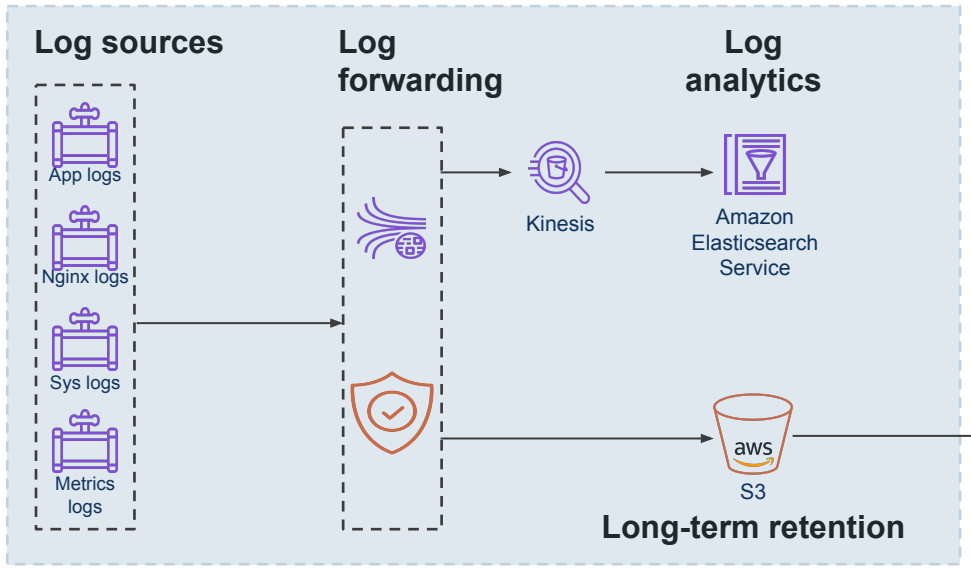
ElasticSearch used for operational analytics, Athena used for ad hoc analytics on logs or BI – both hard to scale

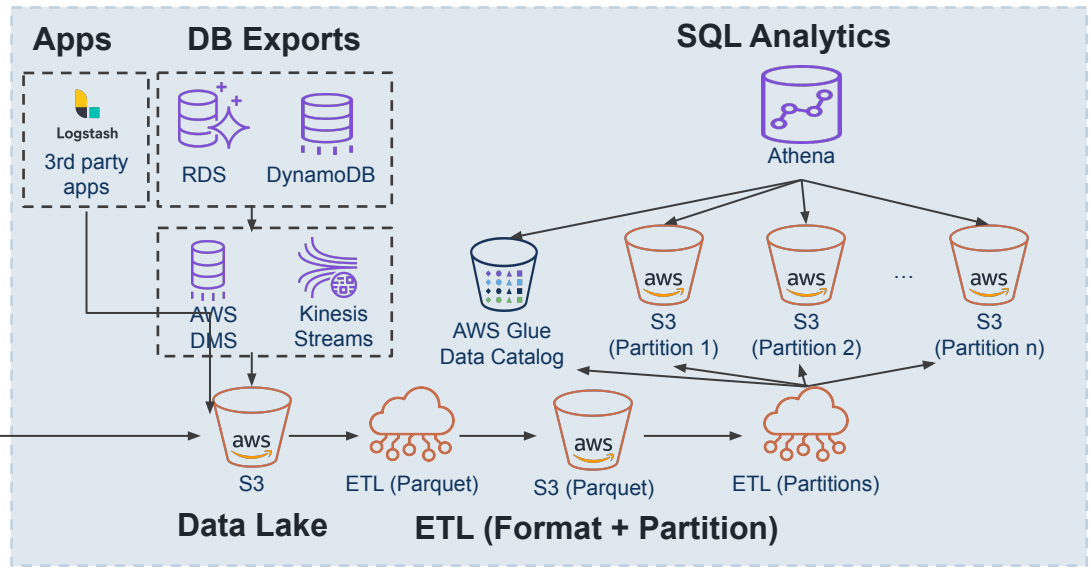
ElasticSearch for operational analytics

- ✔ Monitoring
- ✔ Troubleshooting
- ✔ Threat hunting

SQL for ad hoc analysis, reporting & BI

- ✔ Historical trend analysis
- ✔ Compliance reporting
- ✔ Business analytics

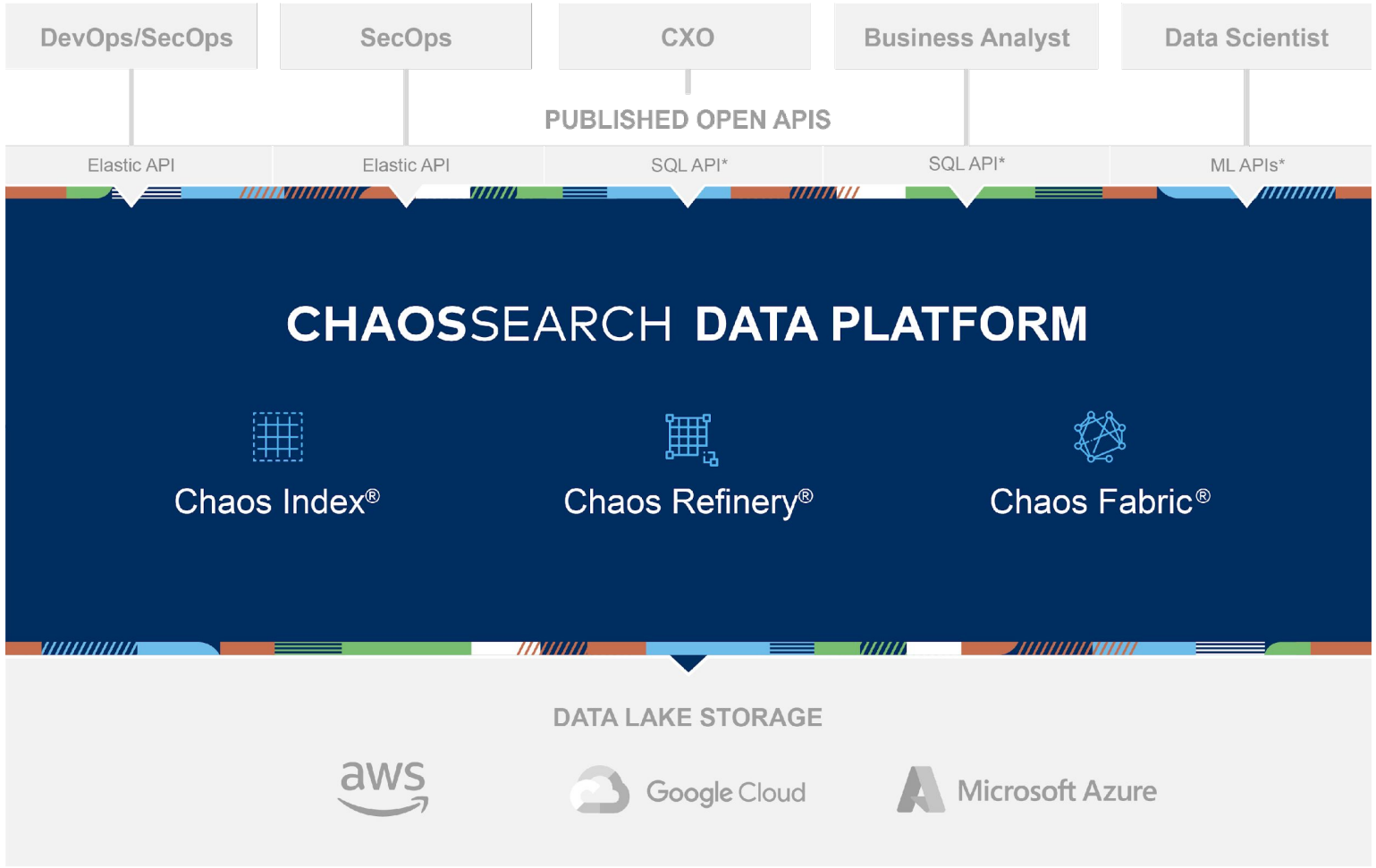




Source: Typical data lake architecture - adapted from “How Affirm leverages AWS to support a unified data lake”

Customers that have Eliminated Complexity with the ChaosSearch Data Lake Platform

ChaosSearch Activates Your Data Lake for Search, SQL and Alerting at Unlimited Scale



Simply ingest raw data and get instant insights out

Flexibility Without Giving Up Performance

✔ Ingest data as-is and query across open analytic APIs

Unlimited Data Retention

✔ No financial tradeoffs that hinder insights

No Data Movement

✔ Simplify your architecture and enhance your security posture

Eliminate Toil and Free Up Resources

✔ Liberate valuable resources from data pipeline creation, constant maintenance and troubleshooting

Superior Cost Economics

✔ Painlessly analyze at petabyte scale while reducing costs by 80%

https://www.chaossearch.io/

Activating Data Lakes for Exploration and Investigation at Unlimited Scale

ChaosSearch enables organizations to Know Better® by performing ad hoc search, SQL and alerting analytics directly against cloud object storage at event log volumes.

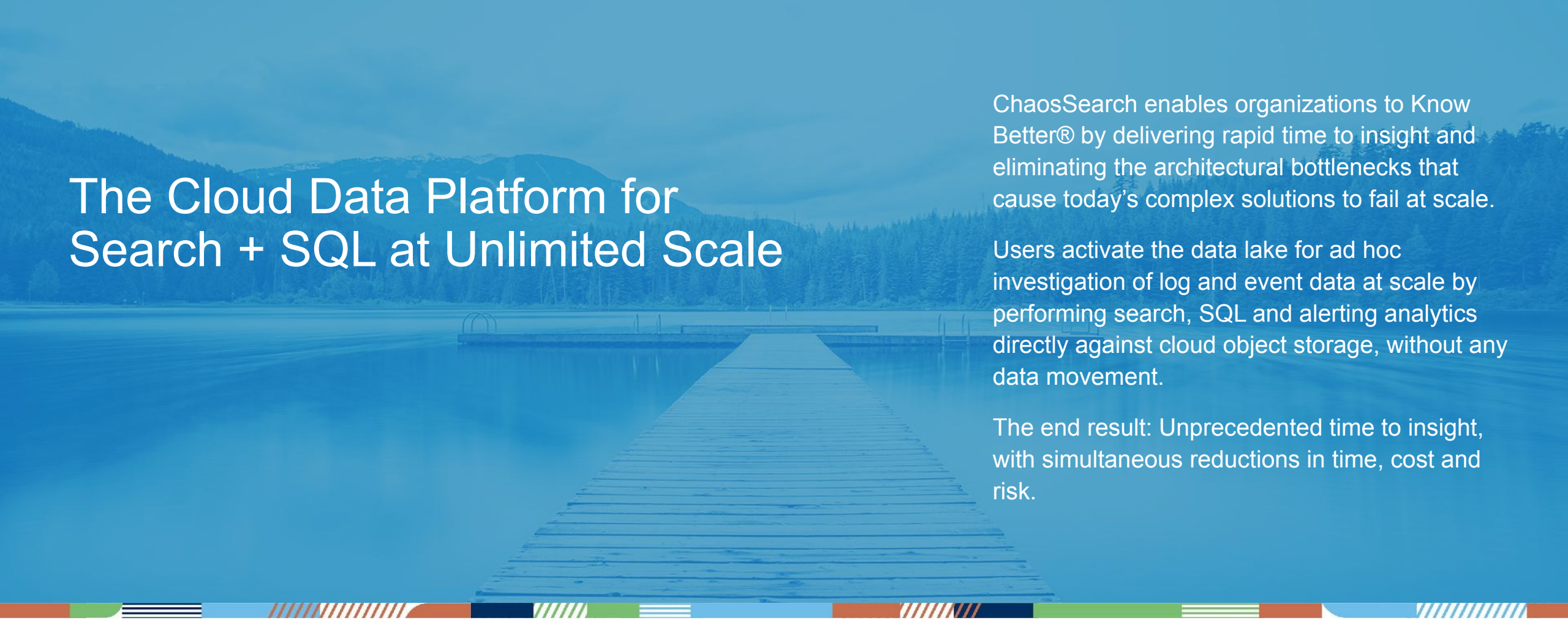
Its Cloud Data Platform delivers unprecedented time to insight by eliminating the architectural bottlenecks and data movement that cause today's complex solutions to fail at scale.

The end result… Simultaneous reductions in time, cost and risk.

FEATURED CUSTOMERS

EQUIFAX · Blackboard · bai communications · REVINATE · Agilence · ARMOR · Klarna. · Digital River

CHAOSSEARCH



The Cloud Data Platform for Search + SQL at Unlimited Scale

ChaosSearch enables organizations to Know Better® by delivering rapid time to insight and eliminating the architectural bottlenecks that cause today's complex solutions to fail at scale.

Users activate the data lake for ad hoc investigation of log and event data at scale by performing search, SQL and alerting analytics directly against cloud object storage, without any data movement.

The end result: Unprecedented time to insight, with simultaneous reductions in time, cost and risk.

FEATURED CUSTOMERS

EQUIFAX · Blackboard · bai communications · REVINATE · Agilence · ARMOR · Klarna. · Digital River

All Current Approaches End in the Same Loop

1. More Boxes
2. More Arrows
3. More Risk
4. More Processes

5. More Complexity

“Optimizations” Add to Architectural Complexity

Specialization

- Each major use case requires their own analytic engine and data schema (Search, SQL, ML)
- Columnar Formats
- Subset
- Shard
- Partition
- Pre-Aggregated Disk Cache
- Memory Cache

“Optimizations” Add to Architectural Complexity

Movement

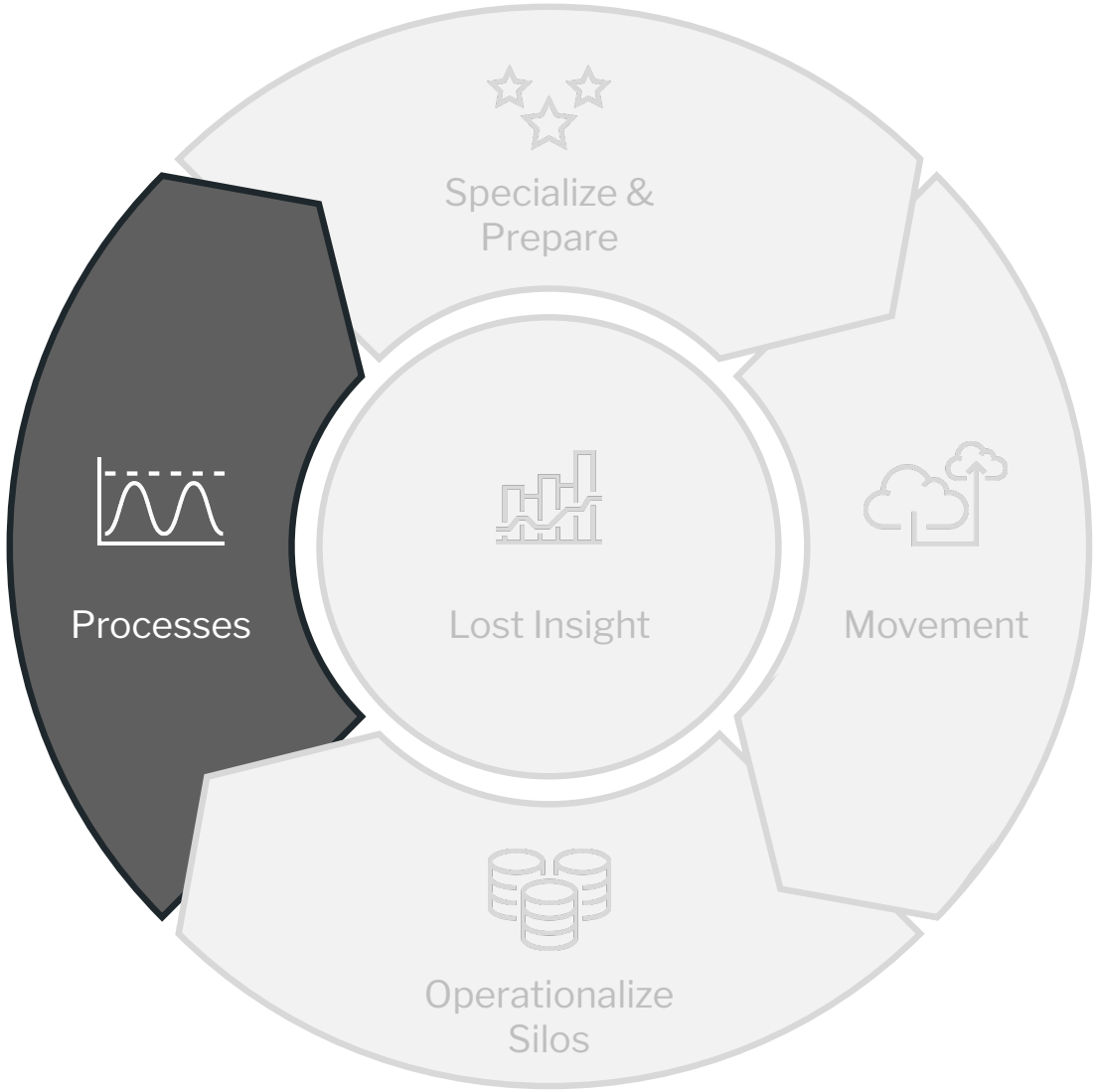
- Complex Data Transformation Pipelines
- Fixed, Single Purpose Schema
- Disk… usually expensive SSDs
- Memory… even more expensive
- With all the copies of data, the source of truth becomes opaque
- Each is biased to the solution build for and gets further from the raw truth

“Optimizations” Add to Architectural Complexity

Operationalization

- Web of pipelines creates significant points of failure and new bugs
- Scale and capacity planning are big tasks
- Capacity planning on “Black Friday” style events cause over provisioning
- More governance and security efforts

“Optimizations” Add to Architectural Complexity



Monitoring

- Continuous Operational Monitoring to Mitigate Risk
- Planning and Rework to Address Every New Issue

Chaotic Process Results in a Loss of Insights

- All while the business is waiting for the right data, in the right place, at the right time to generate value
- Gaps in data access, time to access and loss of insights
- And worse… loop starts all over again with the next question asked at scale!
- Systems and processes “collapse under their own weight”

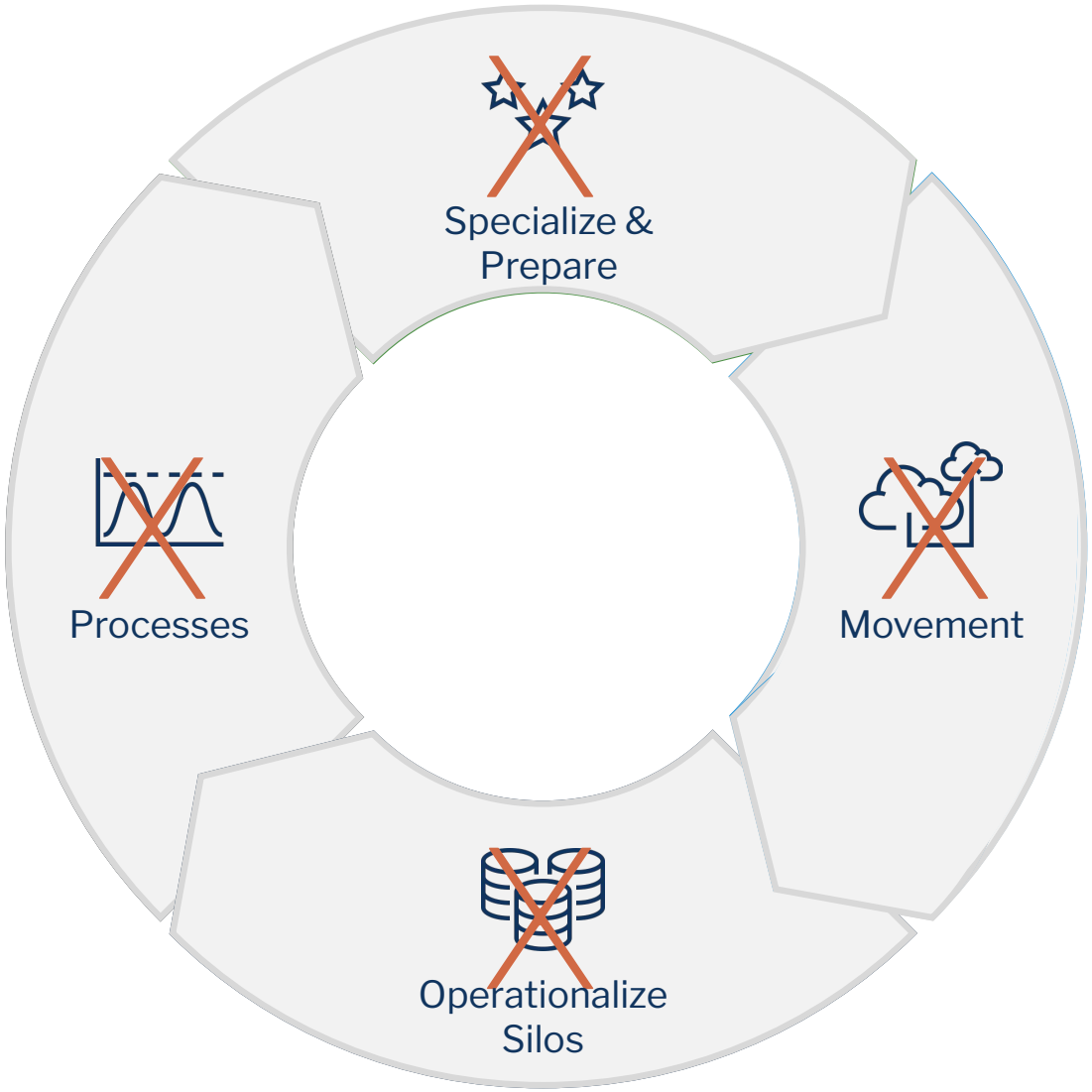
All Current Approaches End in the Same Loop

- All other cloud databases, warehouses and lakehouses are overly dependent on data movement into single-purpose, partitioned structures (usually columnar)
- With complex SSD persistence and/or transient in-memory caches.
  - These caches are Tightly Coupled Storage + Compute!
  - Which is the same architecture as Traditional!
- Result is
  - performance OR scale
  - at the expense of linear direct costs
  - and tremendous complexity.

“With the move to ChaosSearch, 98% of all operational burdens have been lifted from us, allowing us to focus on Anthology-specific tasks.”

Joel Snook, Director, DevOps Engineering

Blackboard
NOW PART OF ANTHOLOGY



Specialization – Eliminated

- ChaosSearch unique data representation enables Schema on Read and supports multiple use case (Search, SQL, ML).
- Additionally, Operational Data (machine generated data) at scale is easily accessible to Business Analytic users with their existing tools, which is typically not the case

Movement - Eliminated

- No data movement is needed with ChaosSearch, just stream your data to your standard cloud storage

Silos - Eliminated

- One unified Data Lake – single source of truth.
- One data representation that has not been manipulated to fit a pre-determined schema per use case.
- Simultaneously support multiple data consumers (Log Analytics, Business Intelligence, Machine Learning, etc.

Processes - Dramatically Streamlined

- No pipelines to create, crash or fall behind during peak loads.
- No cluster or schema to manage.
- Queries leverage dynamic compute pool, no need for capacity planning or over provisioning