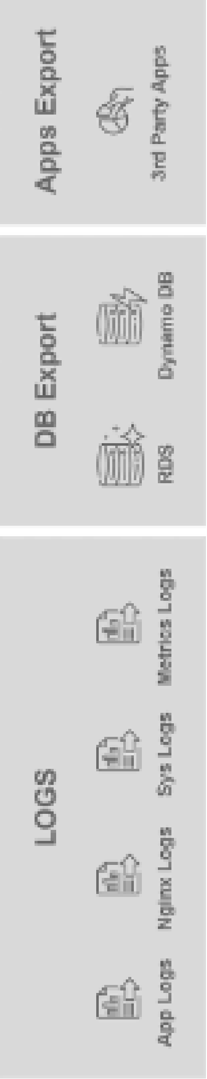
Activating Data Lakes for Analytics at Scale

CHAOSSEARCH

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

& Dave Armlin

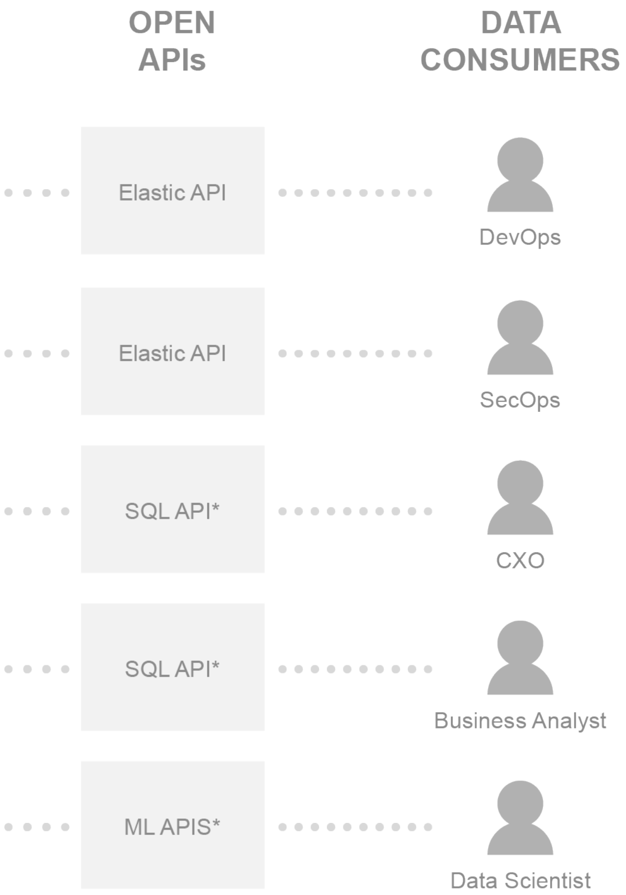


"80% of analytic workers' time is spent getting the right data, to the right place, at the right time".

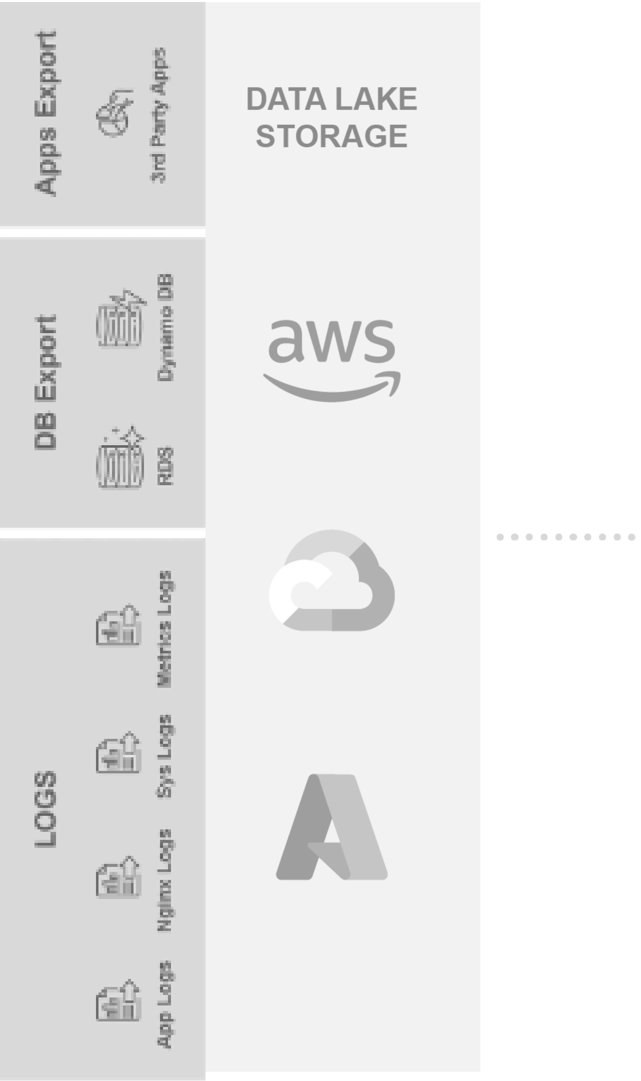
- 20% searching for right data
- 37% preparing into right place and format
- 24% protecting and governing
- Using 4 to 7 different tools, adding to the complexity
- With 44% of workday spent on <u>unsuccessful</u> data activities

Source: IDC

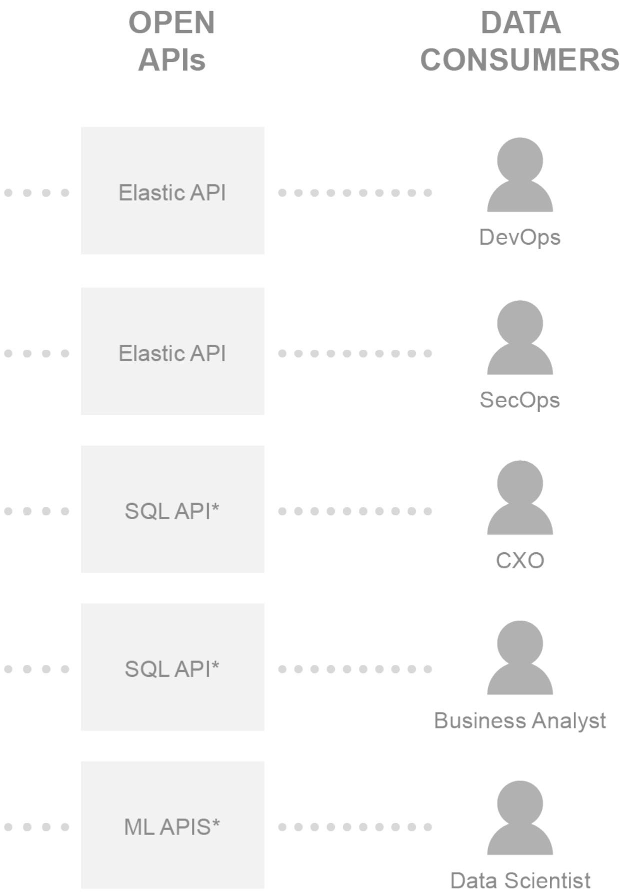
Source: IDC



RAW DATA → **MULTIPLE MODES OF ANALYTICS**

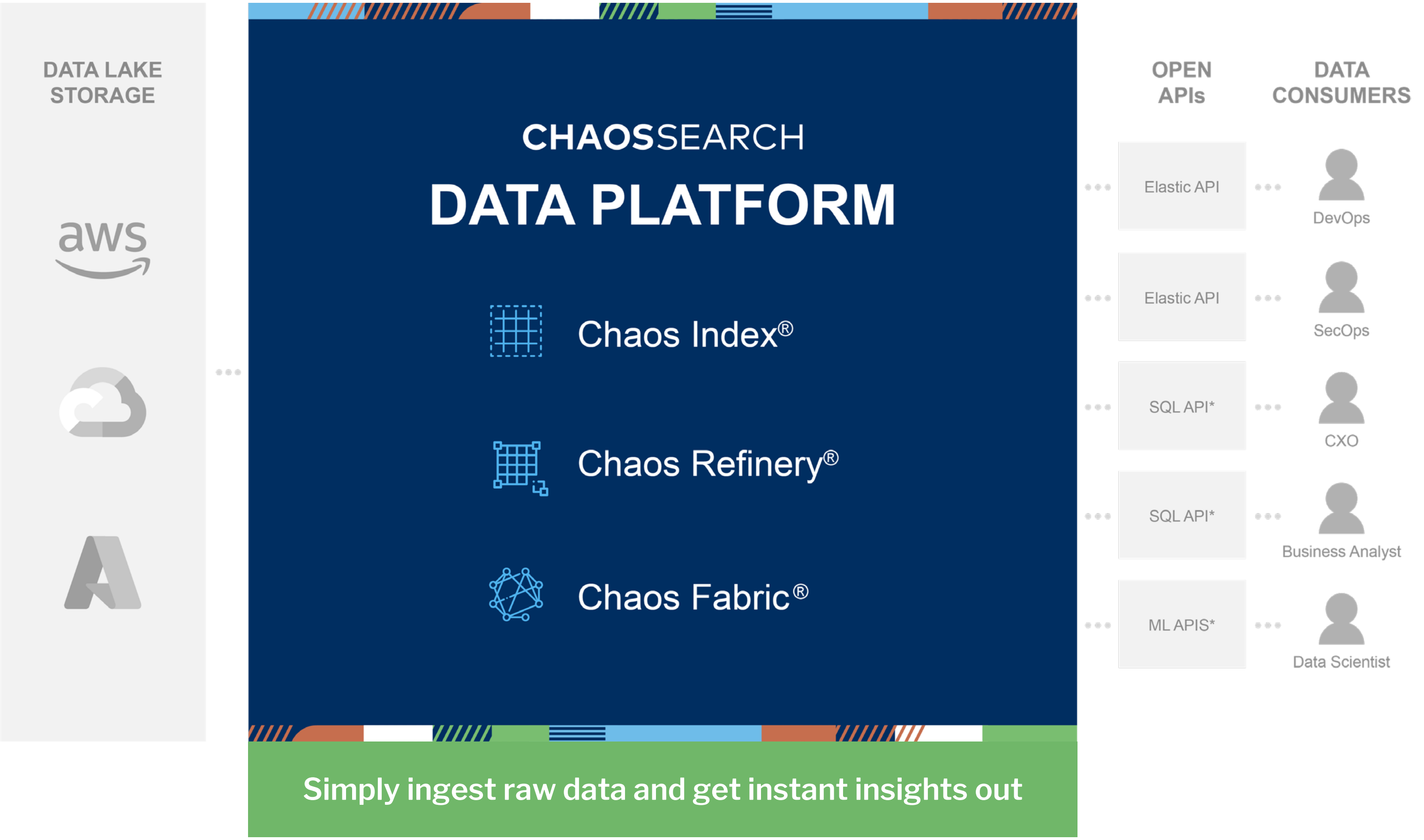


ChaosSearch was founded to solve the challenges of getting from a raw data lake to analytics at scale



RAW DATA → **MULTIPLE MODES OF ANALYTICS**

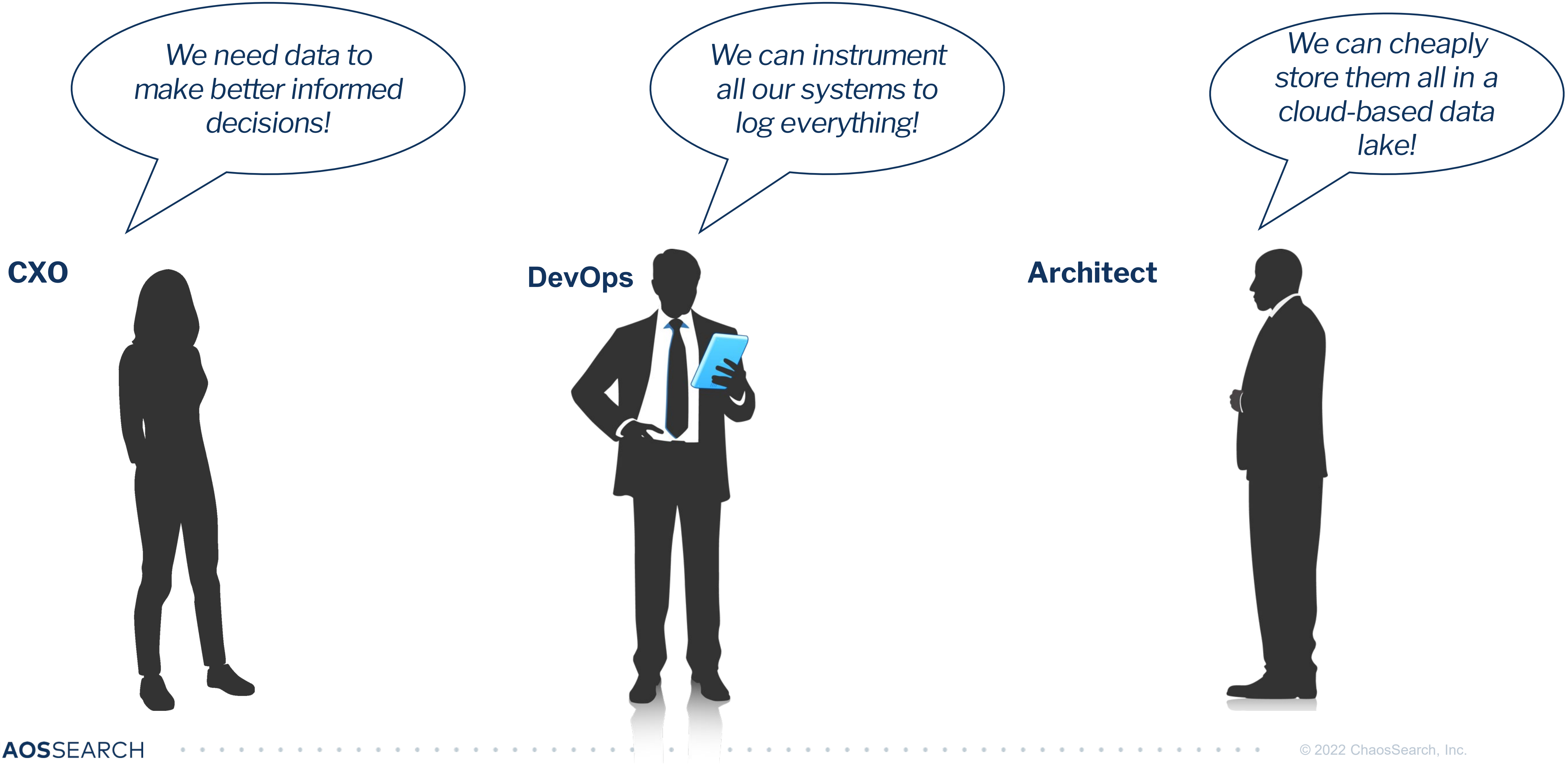
ChaosSearch Activates Your Data Lake for Search, SQL and Alerting at Unlimited Scale



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.**

Operational Analytics Start Simple...



CXO

We need data to make better informed decisions!

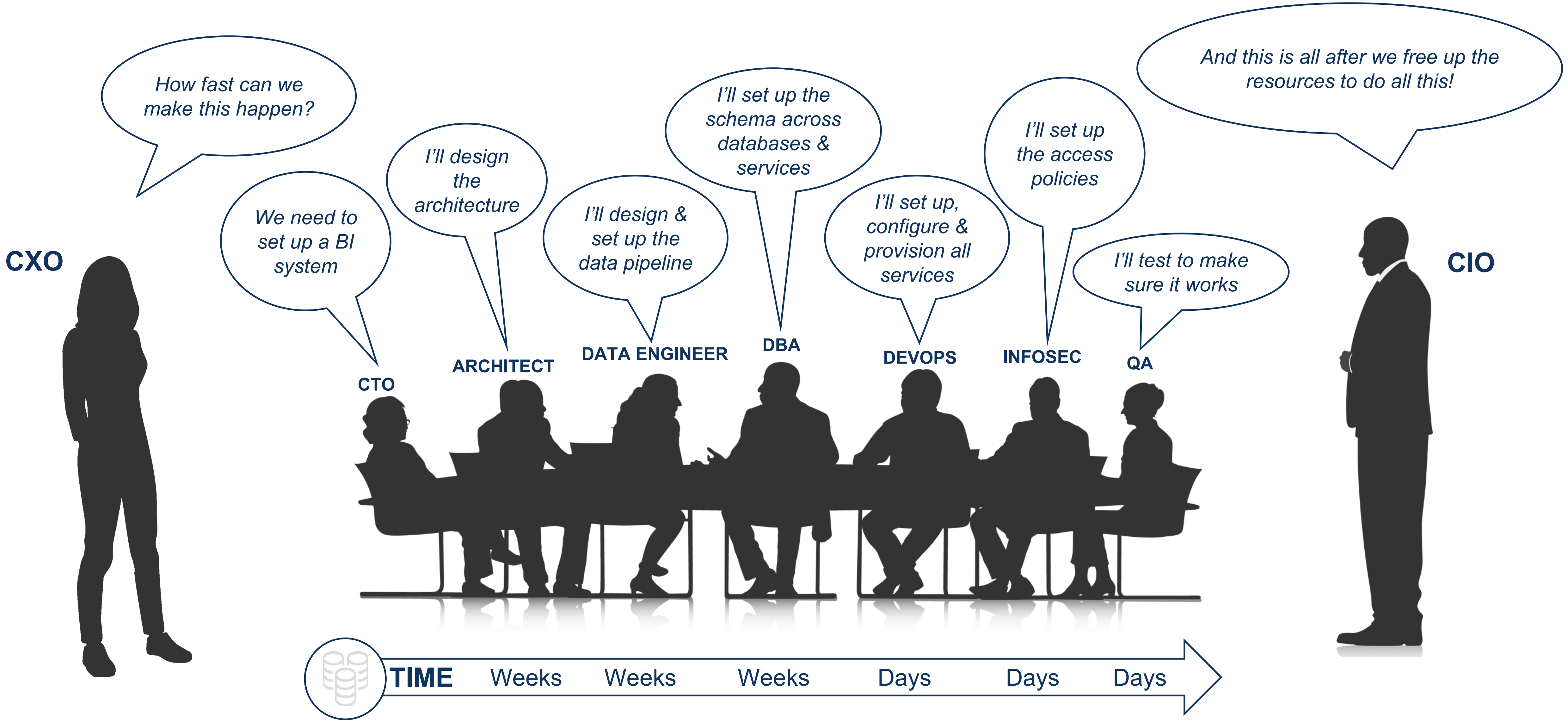
DevOps

We can instrument all our systems to log everything!

Architect

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

But Quickly Get Complicated



CXO: How fast can we make this happen?

CTO: We need to set up a BI system

ARCHITECT: I'll design the architecture

DATA ENGINEER: I'll design & set up the data pipeline

DBA: I'll set up the schema across databases & services

DEVOPS: I'll set up, configure & provision all services

INFOSEC: I'll set up the access policies

QA: I'll test to make sure it works

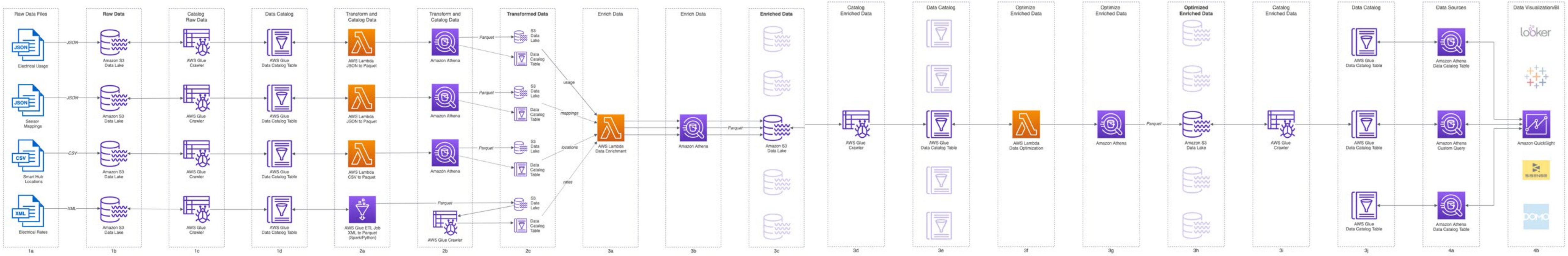
CIO: And this is all after we free up the resources to do all this!

TIME Weeks Weeks Weeks Days Days Days

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

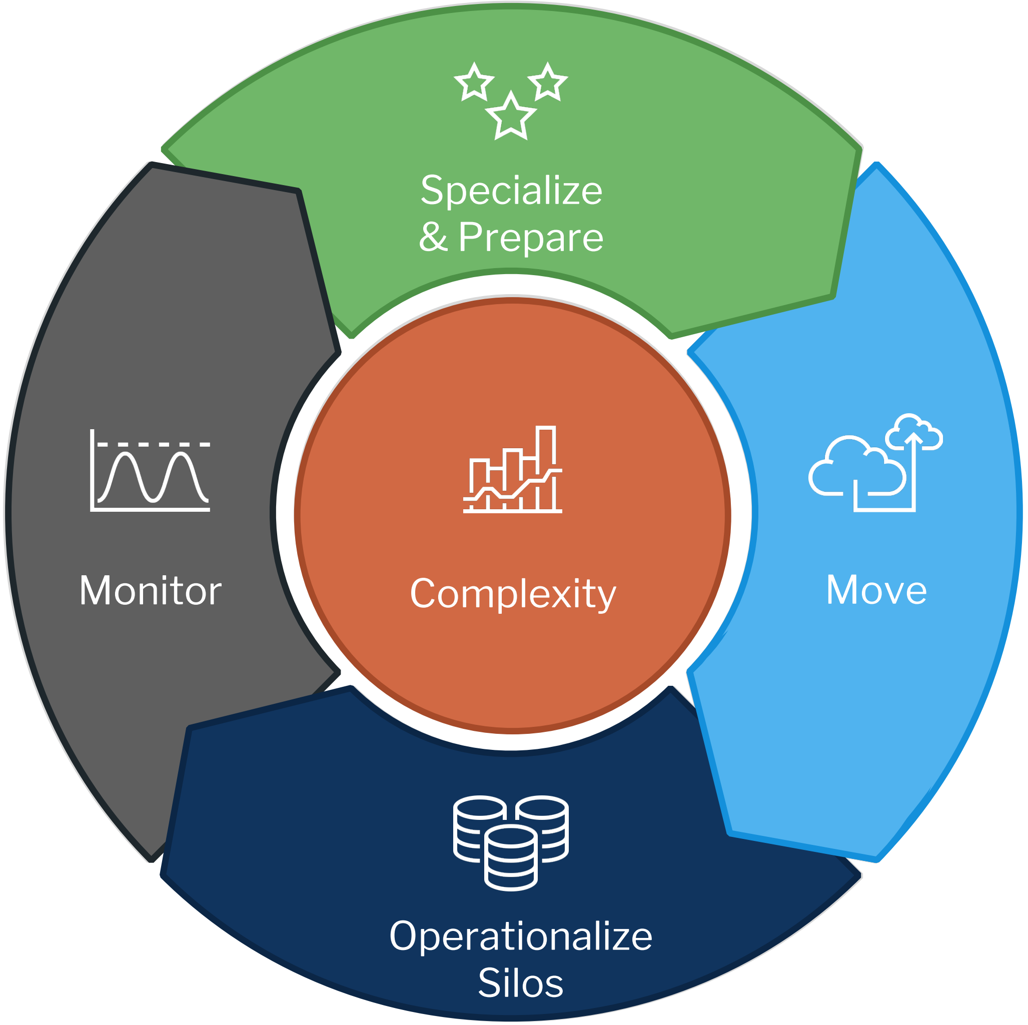
But Scale is the Breaking Point…
With Today’s Cures Forcing a Trade Off

- 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

Less Insights =
Less Value

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

STORE

Connect to any and all data in **your existing** cloud object storage

INDEX

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

REFINE

Prepare your data views for governance & analytics …with no data movement

ANALYZE

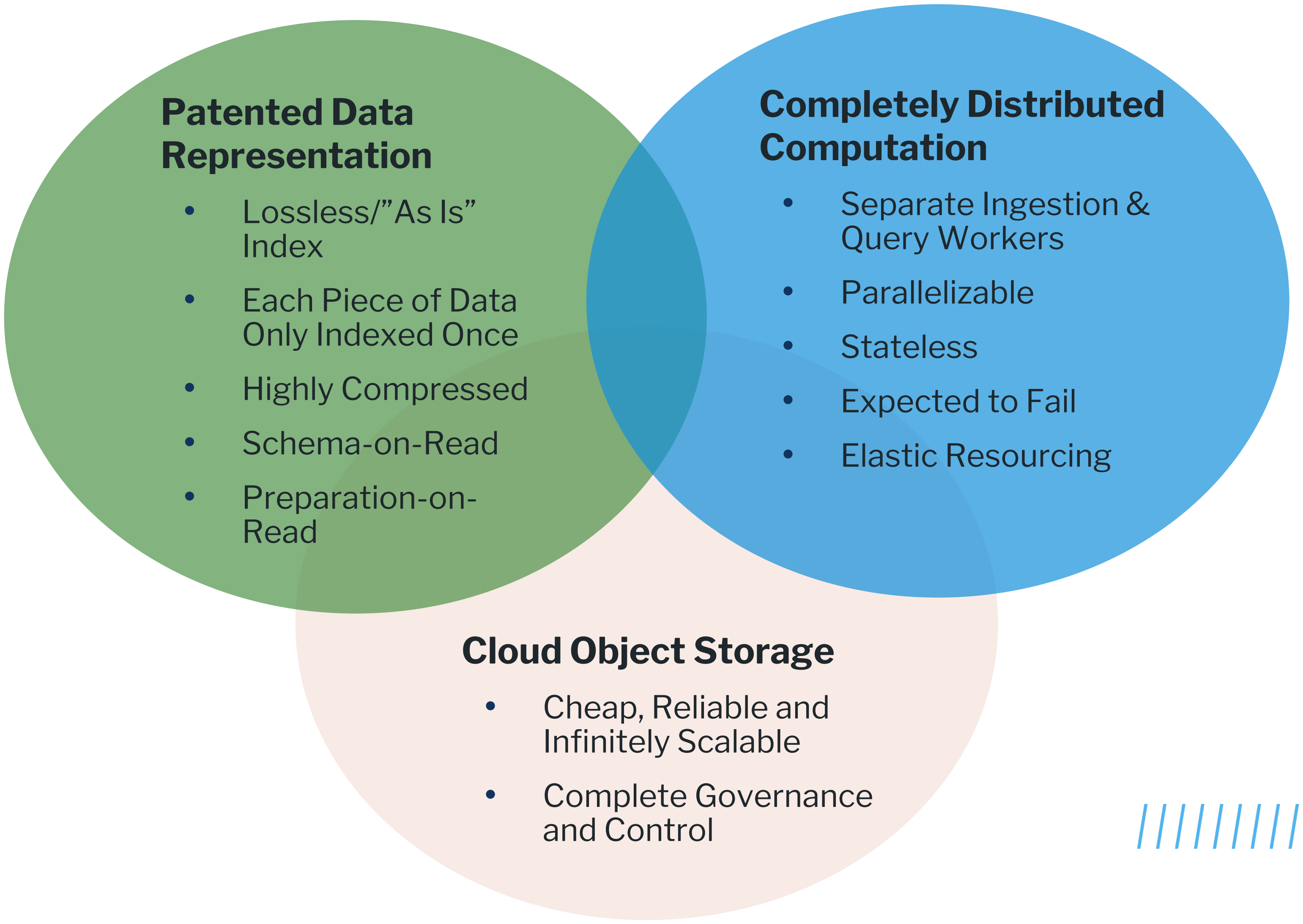
Use your tool of choice for

- Log Analytics
- Exploratory BI
- Continuous Metric 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



Patented Data Representation

- Lossless/"As Is" Index
- Each Piece of Data Only Indexed Once
- Highly Compressed
- Schema-on-Read
- Preparation-on-Read

Completely Distributed Computation

- Separate Ingestion & Query Workers
- Parallelizable
- Stateless
- Expected to Fail
- Elastic Resourcing

Cloud Object Storage

- Cheap, Reliable and Infinitely Scalable
- Complete Governance and Control

- 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

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

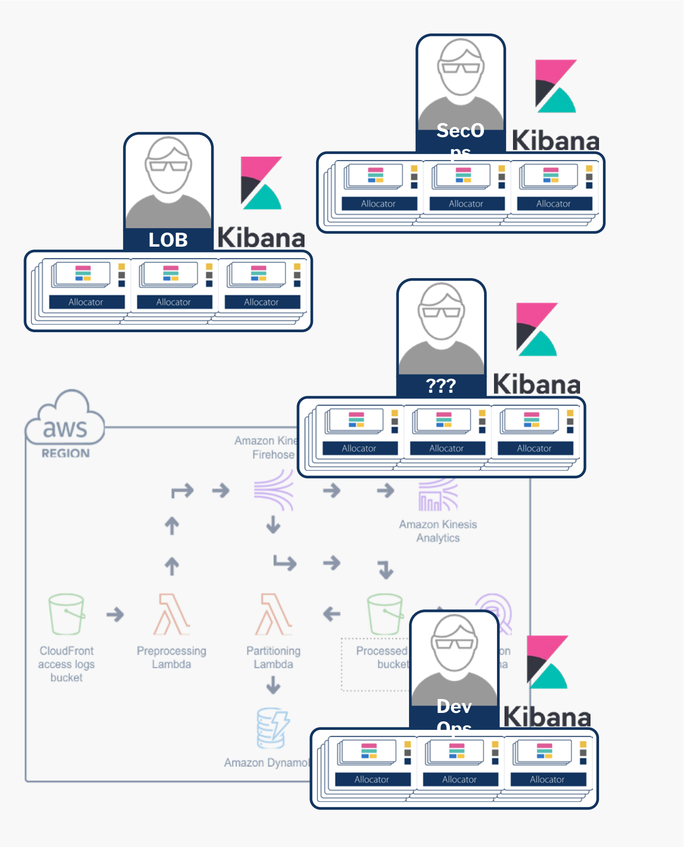
SecOps

- 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

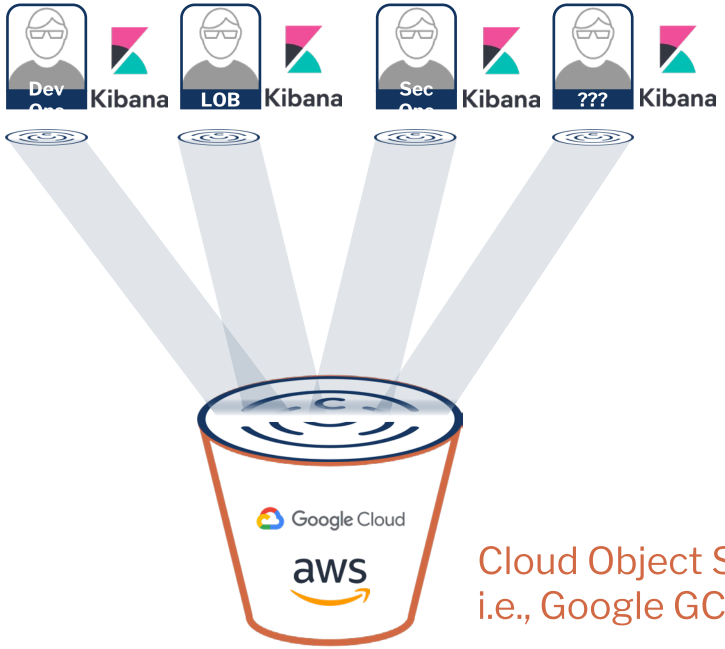
Replacing Elasticsearch or AWS OpenSearch for log analytics at scale

Before: Elasticsearch (ELK stack)



- Limited retention
- Expensive to scale
- Management and configuration challenges
- Downtime created by instability at scale
- Multiple data silos created due to the limits above

With ChaosSearch



PUBLISHED
ELASTIC API

Cloud Object Storage
i.e., Google GCS, AWS S3

One unified data lake

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

Search + SQL pervasive today, but siloed & not built for scale

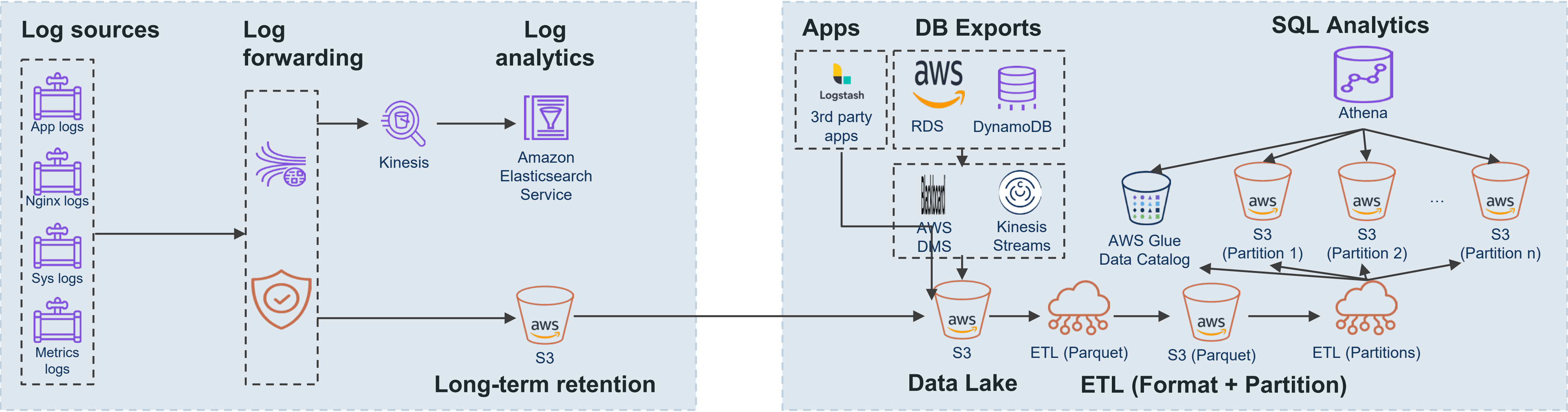
ElasticSearch used for operational analytics, AWS Athena used for ad hoc analytics on logs or BI – both hard to scale

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

RIPPLING

bai communications

REVINATE

Agilence

Blackboard

SIXTH STREET

modicagroup

instruct ERIC

Obelis GROUP

orgvue

TRANSEO

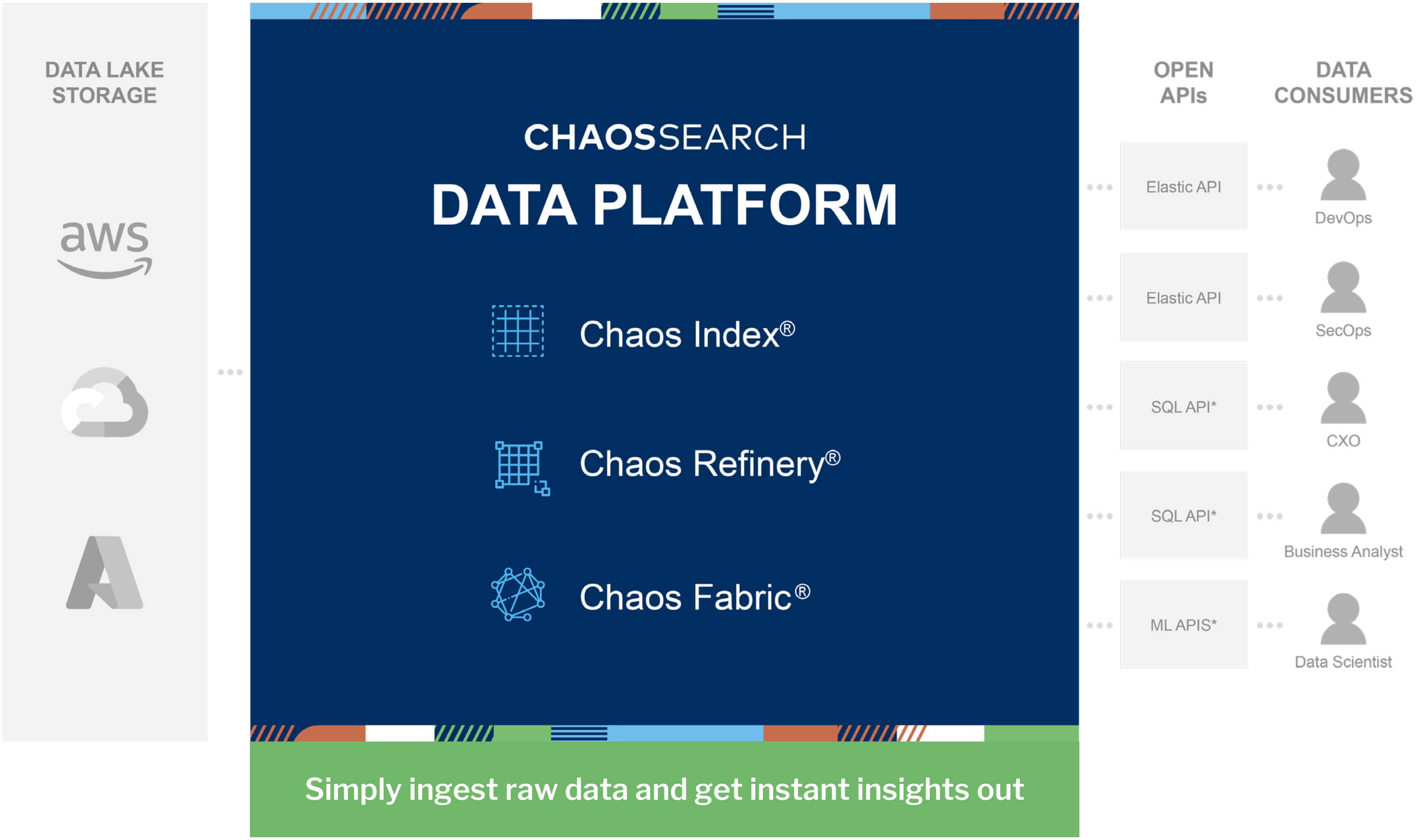
Klarna.

Digital River

EQUIFAX

ARMOR

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



DATA LAKE STORAGE

aws

CHAOSSEARCH

DATA PLATFORM

Chaos Index®

Chaos Refinery®

Chaos Fabric®

Simply ingest raw data and get instant insights out

OPEN APIs

DATA CONSUMERS

Elastic API — DevOps

Elastic API — SecOps

SQL API* — CXO

SQL API* — Business Analyst

ML APIS* — Data Scientist

Unlimited Data Retention

✓ No financial tradeoffs that hinder insights and create vulnerabilities

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/