soda foundation

IEEE COMPUTER SOCIETY

Discovering Unstructured Data!

Sanil KUMAR

TOC Co-Chair, SODA Foundation | Founder CEO, Caze Labs
LinkedIn: sanilkumard

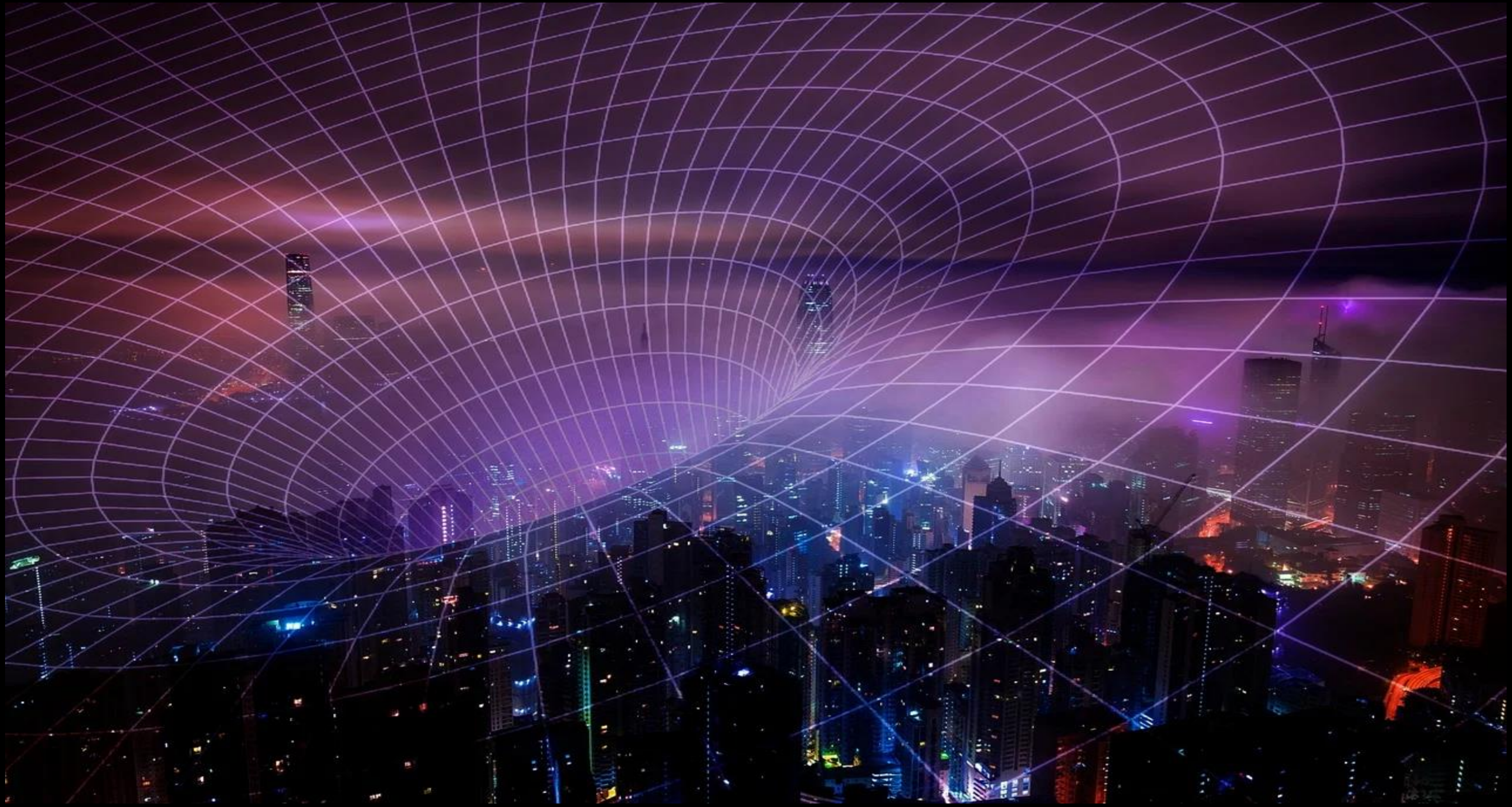
“Data First, What Next?”

Systems & Software Conclave, IEEE Computer Science Week 2023

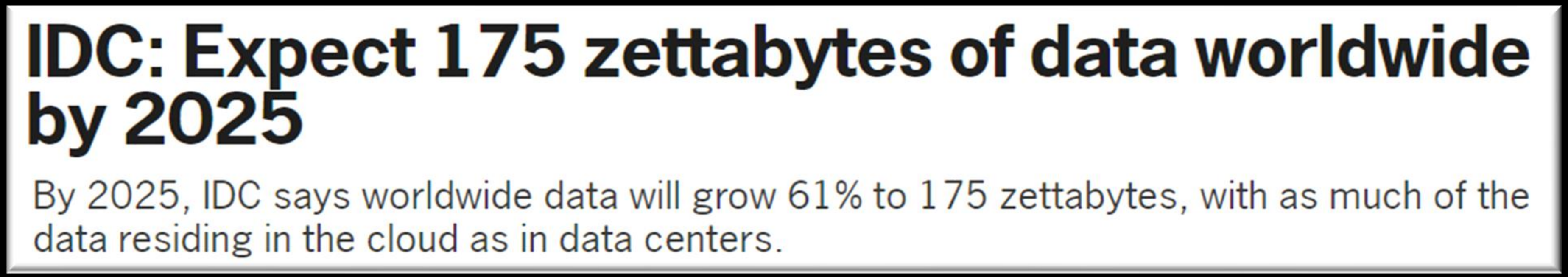
In Collaboration with SODA Foundation

30th Nov 2023, Ericsson, Bangalore

IMMERSED IN DATA...



EXPLOSION OF DATA…



IDC: Expect 175 zettabytes of data worldwide by 2025

By 2025, IDC says worldwide data will grow 61% to 175 zettabytes, with as much of the data residing in the cloud as in data centers.

2023 : **120 ZB**

2025 : **181 ZB**

https://www.statista.com/

1 ZB = 1000^7 bytes = 10^{21} bytes = 1000000000000000000000 bytes = 1000 exabytes

DYNAMIC TECHNOLOGY CONTEXT!

80%

Adopt Multicloud
Leverage Multicloud Strategy

DYNAMIC TECHNOLOGY CONTEXT!

LLM

Huge impact of Generative AI
GPT-3 and GPT-4

Vector Databases

Data Mesh

DYNAMIC TECHNOLOGY CONTEXT!

125%

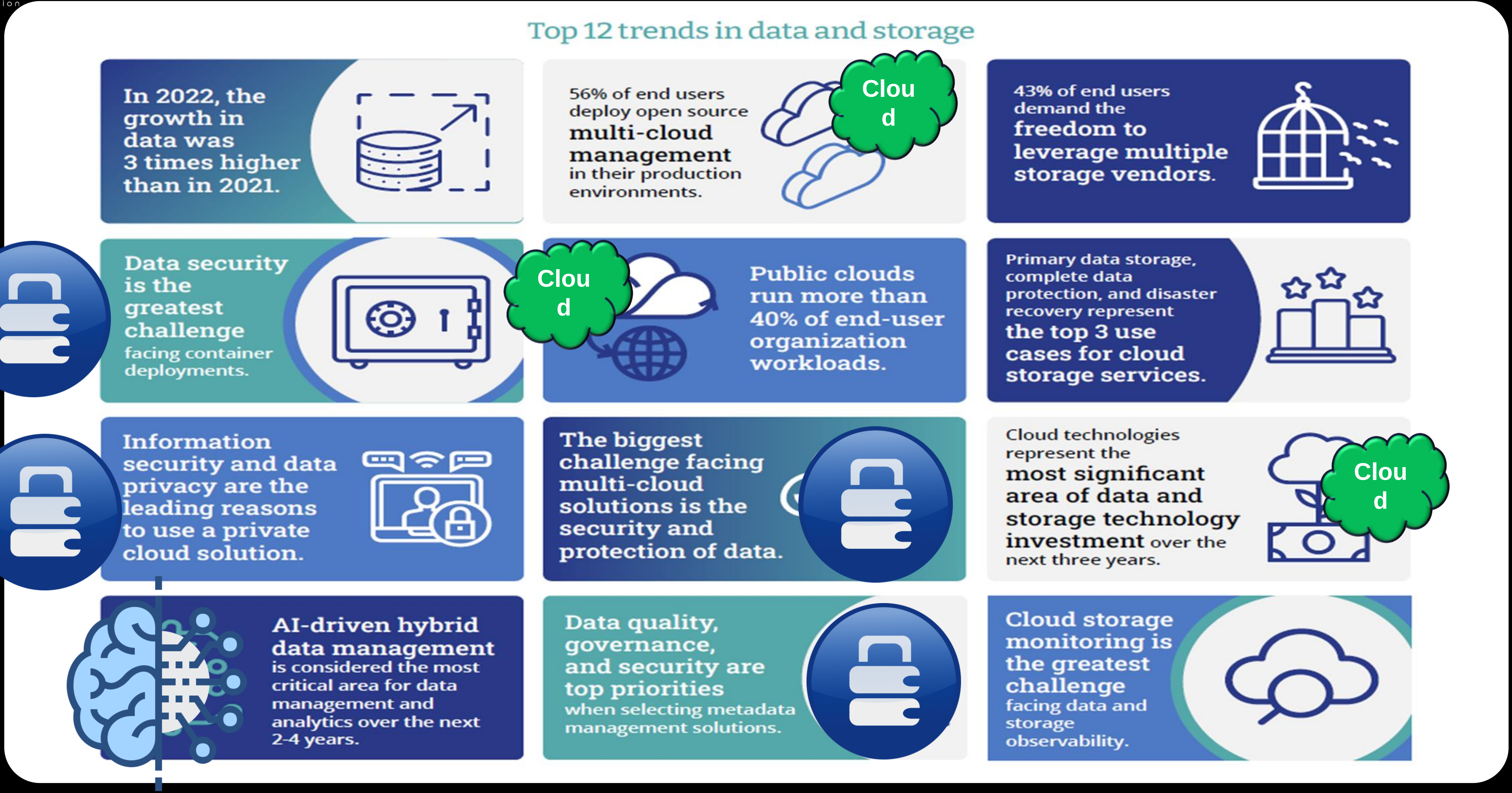
Cyber Security Attacks increased

$4.35M

Average Data Breach Cost

39%

Internet users attacked (avg) globally in 2022



Top 12 trends in data and storage

In 2022, the growth in data was 3 times higher than in 2021.

56% of end users deploy open source **multi-cloud management** in their production environments.

Clou
d

43% of end users demand the **freedom to leverage multiple storage vendors.**

Data security is the greatest challenge facing container deployments.

Clou
d

Public clouds run more than 40% of end-user organization workloads.

Primary data storage, complete data protection, and disaster recovery represent **the top 3 use cases for cloud storage services.**

Information security and data privacy are the leading reasons to use a private cloud solution.

The biggest challenge facing multi-cloud solutions is the security and protection of data.

Cloud technologies represent the **most significant area of data and storage technology investment** over the next three years.

Clou
d

AI-driven hybrid data management is considered the most critical area for data management and analytics over the next 2-4 years.

Data quality, governance, and security are top priorities when selecting metadata management solutions.

Cloud storage monitoring is the greatest challenge facing data and storage observability.

Multicloud

- Data across multiple clouds
- Hybrid Data Management
- Cost and QoS

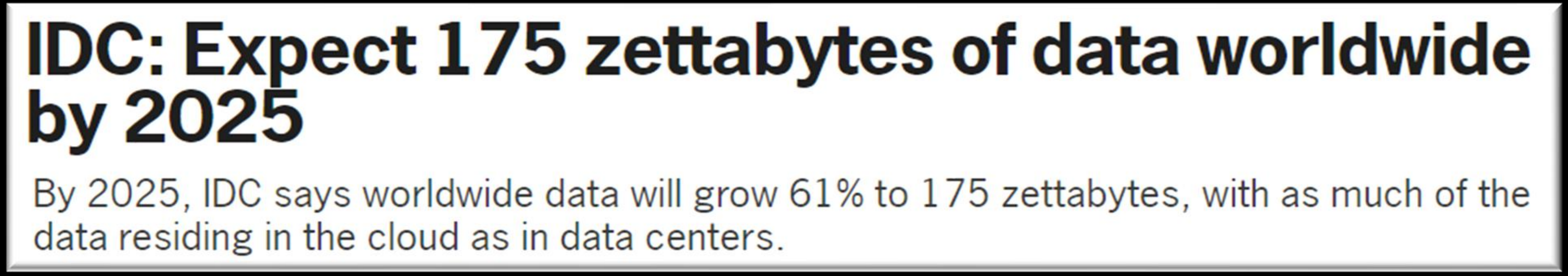
AI

- Analytics and Inference
- Data Management Efficiency
- Visualization

Security

- Data Privacy
- Data Confidentiality
- Data Integrity

EXPLOSION OF UNSTRUCTURED DATA…

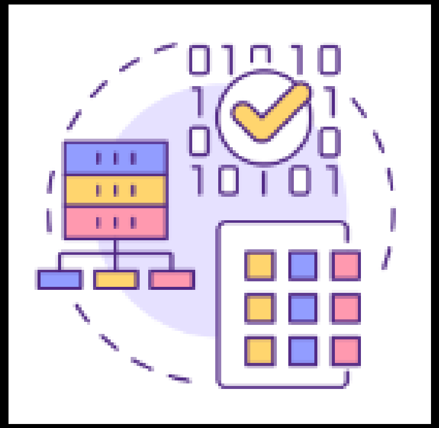
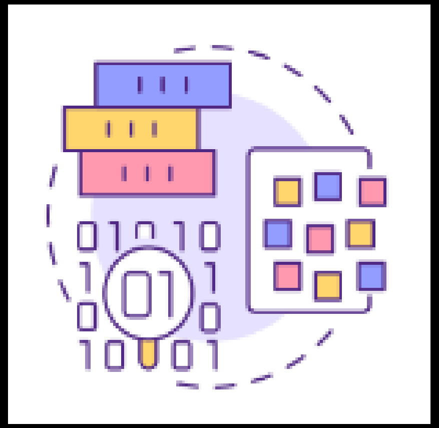


IDC: Expect 175 zettabytes of data worldwide by 2025

By 2025, IDC says worldwide data will grow 61% to 175 zettabytes, with as much of the data residing in the cloud as in data centers.

with 80% of that data being unstructured!

90% of it is NEVER analyzed!

STRUCTURED	UNSTRUCTURED
	
Organized	Not Organized
Formatted	Not Formatted
Schema	No Schema

DARK DATA

Gartner defines dark data as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing).

typically unstructured or semi-structured

kind of invisible, not part of decision making

UNSTRUCTRED DATA MANGEMENT

Challenges

Scattered and Invisible

Privacy Concerns

Not compatible, so lack of tools

Silos

Distributed trend

UNSTRUCTRED DATA MANGEMENT

State forward:

AI / ML

Cloud based Data Management

Data Catalogs and Metadata management

Changing Privacy Situation (Tech for anonymization & differential privacy, compliance, governance)

Edge Computing

Blockchain

Storage advancements (dedup, compression, tiered storage)

...

What are we doing at The SODA Foundation?

Research, Industry Analysis and Crystal

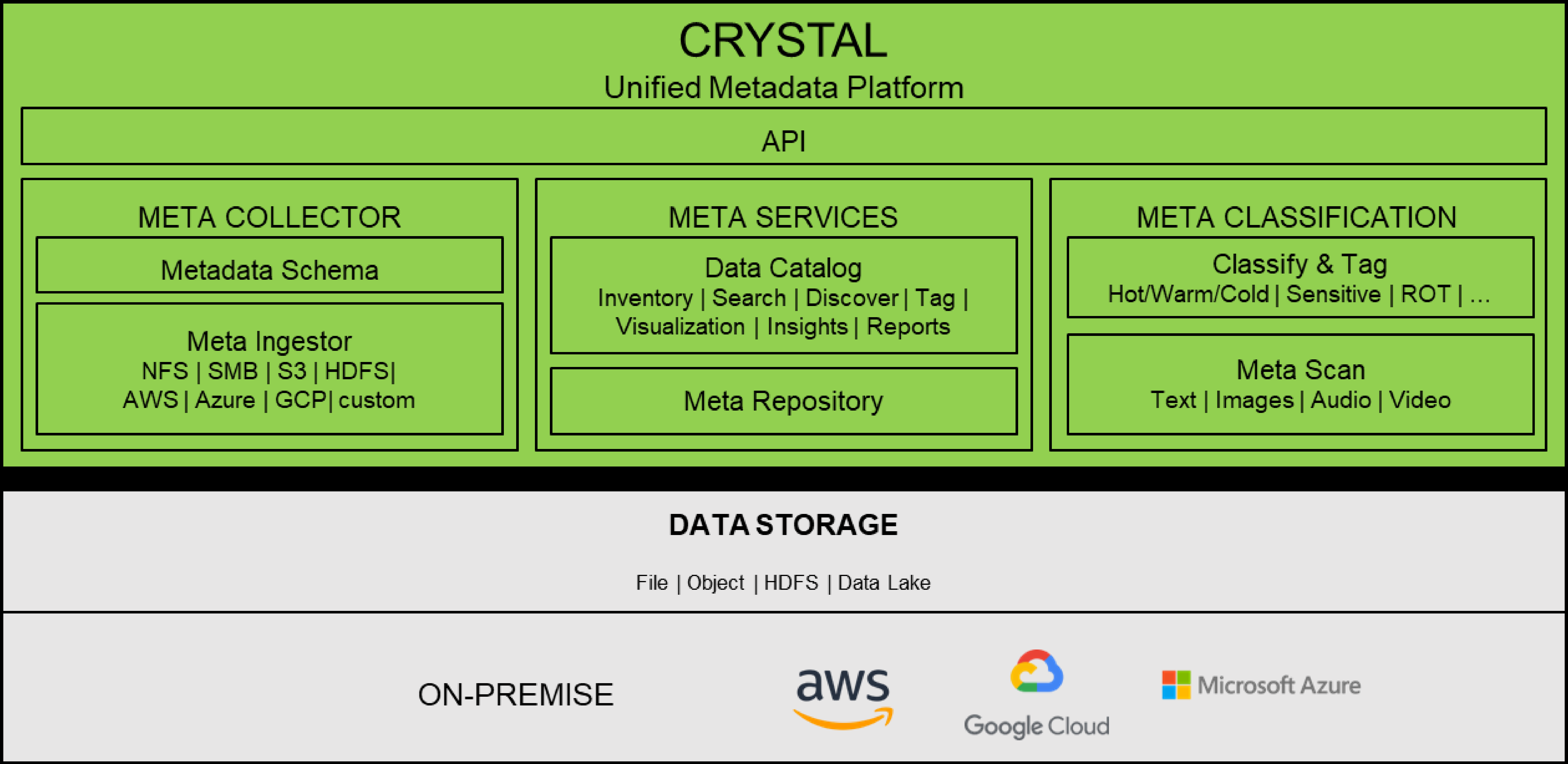
ABSTRACT DATA SILO COMPLEXITY

SIMPLIFY DATA **DISCOVERY**

CRYSTAL

Unified View Of Data

Unified metadata for unstructured data across on-premise and cloud storage



CRYSTAL

Unified Metadata Platform

API

META COLLECTOR

Metadata Schema

Meta Ingestor
NFS | SMB | S3 | HDFS|
AWS | Azure | GCP| custom

META SERVICES

Data Catalog
Inventory | Search | Discover | Tag |
Visualization | Insights | Reports

Meta Repository

META CLASSIFICATION

Classify & Tag
Hot/Warm/Cold | Sensitive | ROT | ...

Meta Scan
Text | Images | Audio | Video

DATA STORAGE

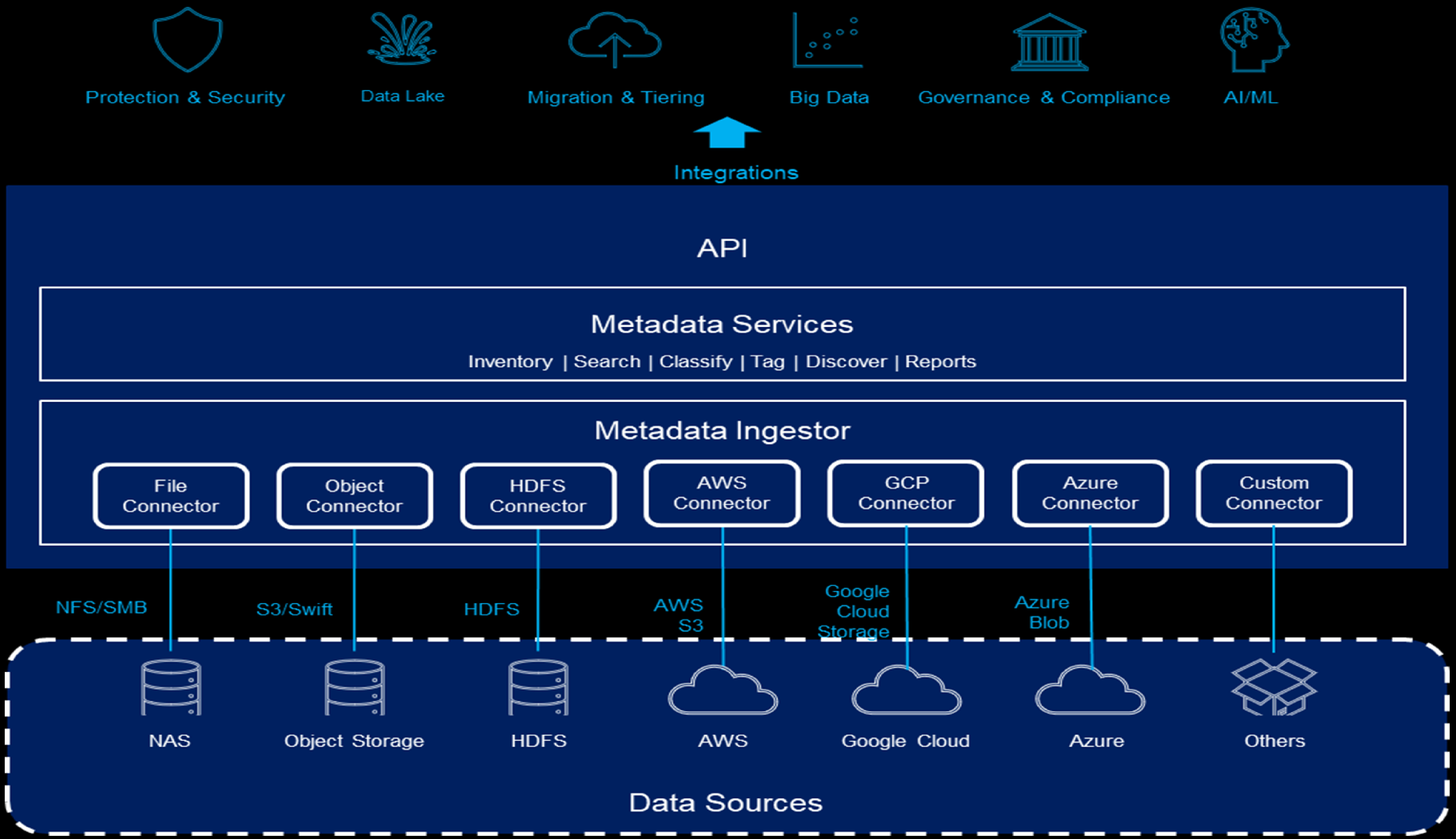
File | Object | HDFS | Data Lake

ON-PREMISE

aws

Google Cloud

Microsoft Azure



Protection & Security

Data Lake

Migration & Tiering

Big Data

Governance & Compliance

AI/ML

Integrations

API

Metadata Services

Inventory | Search | Classify | Tag | Discover | Reports

Metadata Ingestor

File Connector

Object Connector

HDFS Connector

AWS Connector

GCP Connector

Azure Connector

Custom Connector

NFS/SMB

S3/Swift

HDFS

AWS S3

Google Cloud Storage

Azure Blob

NAS

Object Storage

HDFS

AWS

Google Cloud

Azure

Others

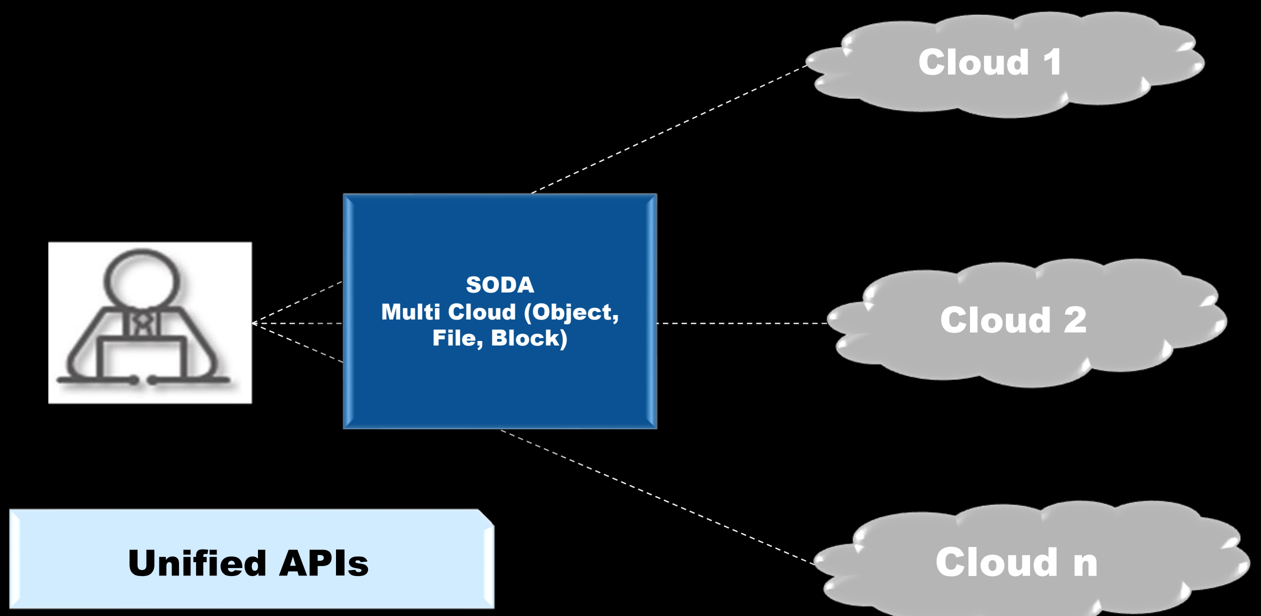
Data Sources

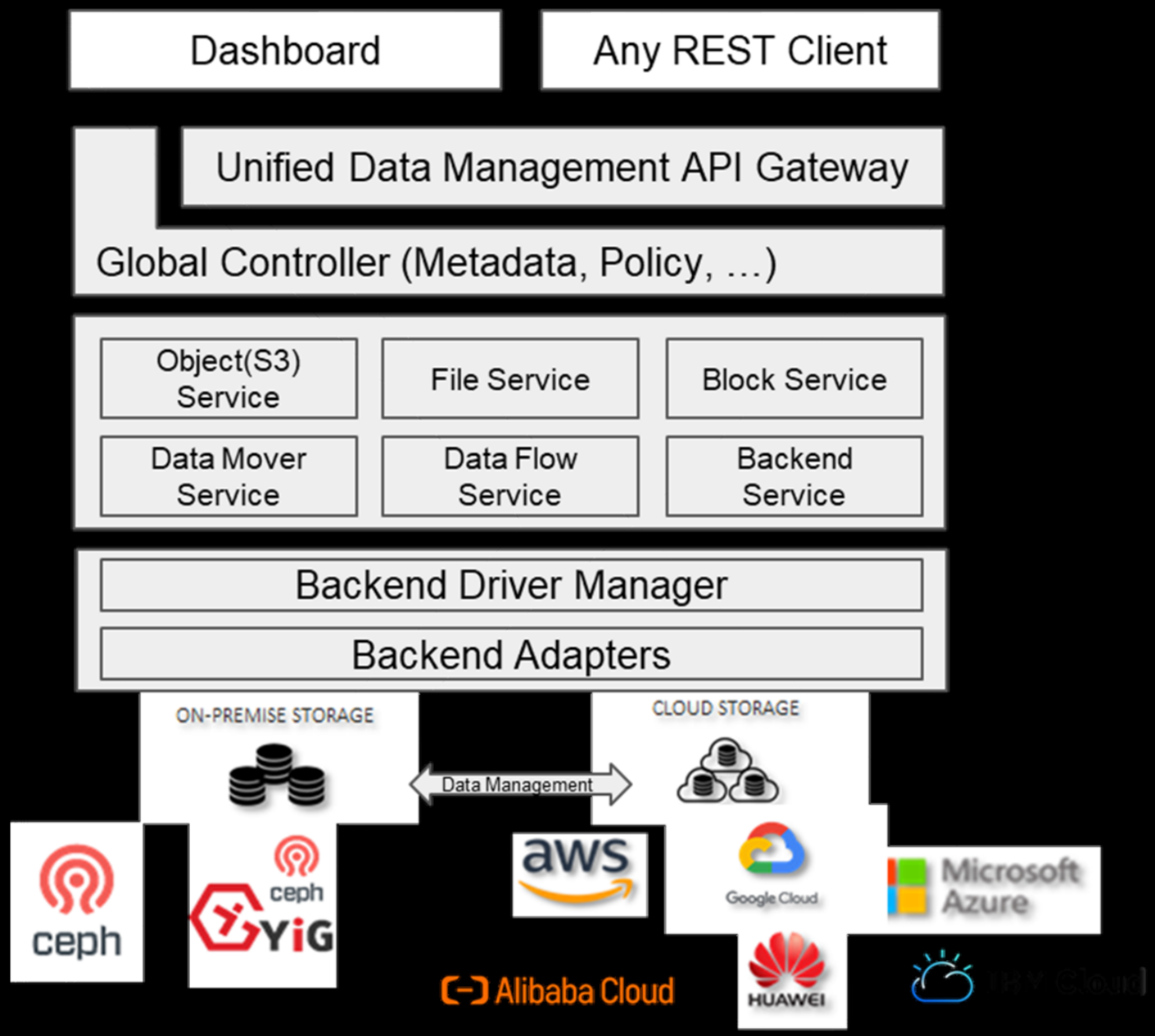
Where are we now?

STRATO

Move data across multicloud environment with a common S3 compatible interface

MOVE




Dashboard

Any REST Client

Unified Data Management API Gateway

Global Controller (Metadata, Policy, …)

Object(S3) Service

File Service

Block Service

Data Mover Service

Data Flow Service

Backend Service

Backend Driver Manager

Backend Adapters

- Unified multicloud API
- S3 compatible
- Support all major cloud vendors
- Data migration
- Data Lifecycle management

- Hybrid data management for object storage (on-prem-cloud)
- File/Block/Object
- User Level Tiering Plan – Storage Service plan
- Smart Archival

https://github.com/sodafoundation/strato

Crystal getting cut!

Crystal Project just started deriving s3 metadata management from SODA Strato

https://github.com/sodafoundation/crystal

Plan Next…

Technology and Project Analysis, Gap, Requirements

Architecture, Design

Crystal first cut Q2 2024

Together we can

shape Open Source Crystal!

discover unstructured data!

Give your requirements and inputs here!
https://github.com/sodafoundation/crystal/issues/9

http://bit.ly/soda-starter

https://www.sodafoundation.io/slack/

Borcelle

THANK YOU

https://bit.ly/soda-starter

www.sodafoundation.io

www.computer.org

soda foundation

IEEE COMPUTER SOCIETY