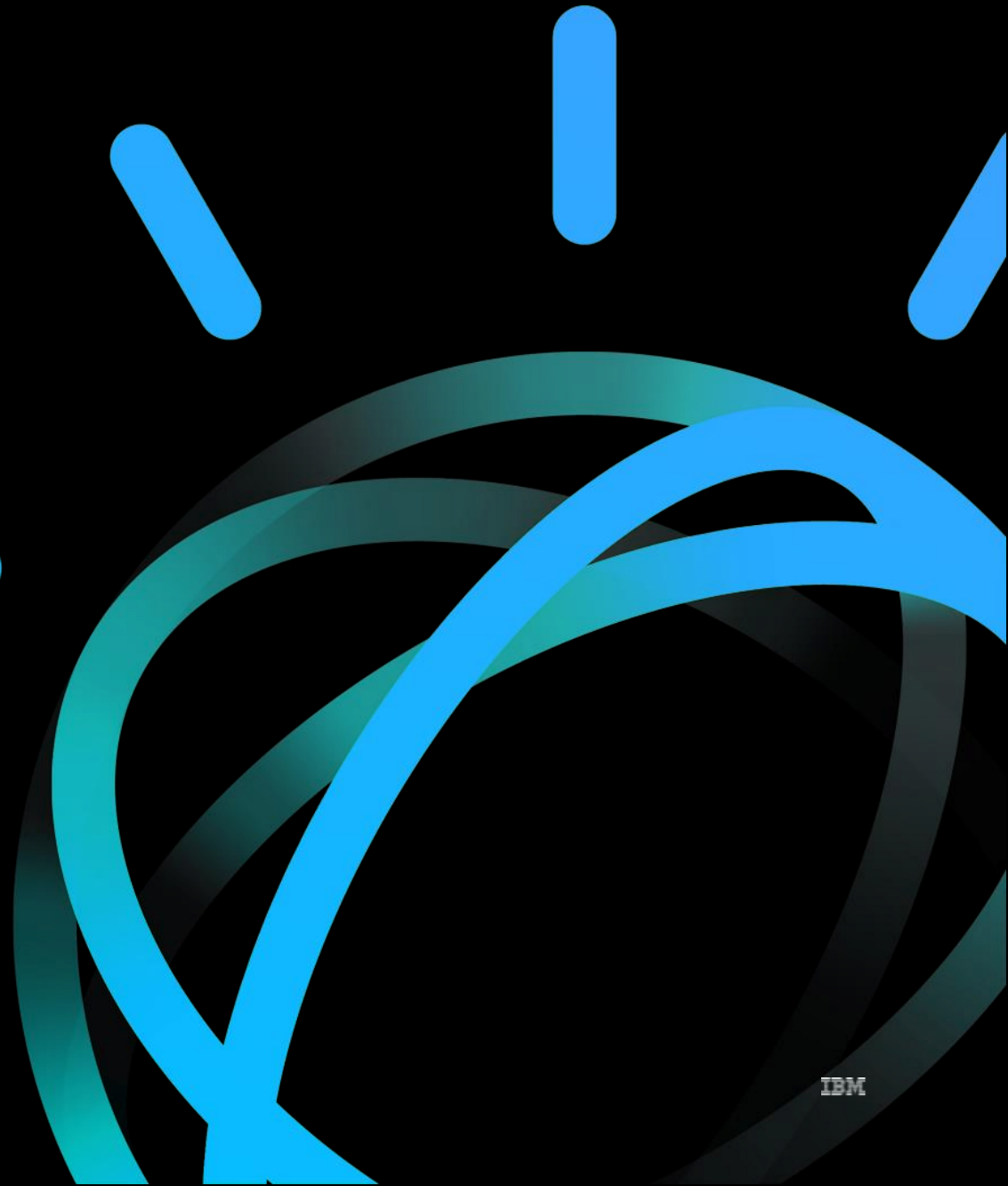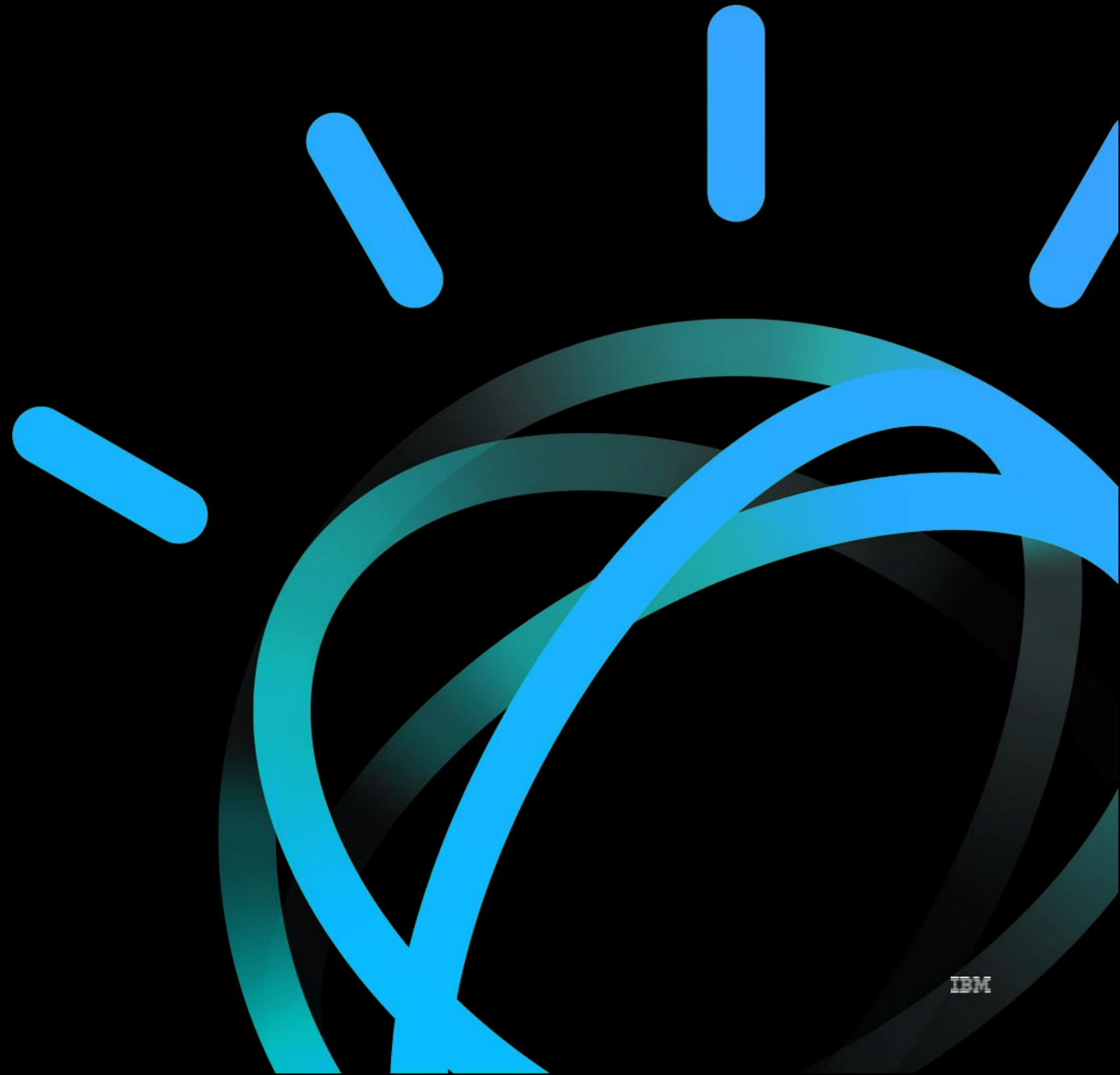# Do you know where your AI data flow bottlenecks are?

IBM

Steven Eliuk, PhD,
VP Deep Learning, IBM GCDO
steven.eliuk@ibm.com

Douglas O'flaherty,
Spectrum Solutions (aka storage)
douglasof@us.ibm.com

There is no AI without an IA

(Information Architecture)

80% of data is either inaccessible, untrusted or unanalyzed

90% say improving the use of data is a top priority

FORRESTER

"No amount of AI algorithmic sophistication will overcome a lack of data [architecture] ... bad data is simply paralyzing

MITSloan
Management Review

The AI Ladder

*A prescriptive approach to accelerating your journey to AI*

INFUSE – Automate and scale across your processes

TRUST – Achieve trust and transparency in outcomes

ANALYZE – Scale insights with Machine Learning everywhere

ORGANIZE – Create a trusted analytics foundation

COLLECT – Make data simple and accessible

MODERNIZE
your data estate for an AI
and multicloud world
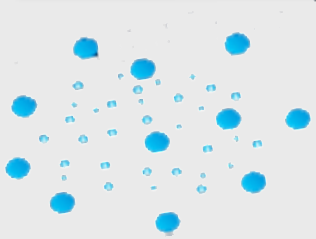
**Data of every type, regardless of
where it lives**

# Data Management Challenges in Analytics and AI

- Data ingest and preparation cycle are too time consuming

- Multi-source data aggregation

- Silos of infrastructure for various analytics use cases

- Multiple copies of same data without a single source of truth

- Analytics on stale data

- Need to securely manage and protect data for traceability

- Need for global accessibility and collaboration

Adopting and Expanding AI

A Single View of the Truth

Provision Time

INGEST
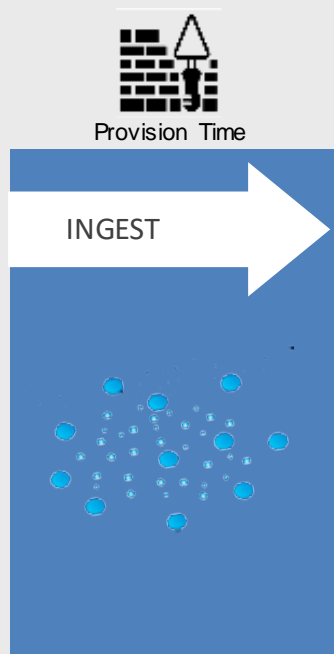
Collect and normalize multiple data sources
- Global requirements: IoT, Mobile, Sensors
- Client data
- Supply Chain Data
- Transactional Systems
- Client Behavior

Standard data analytics tools extract relevant data
- ETL: Extract, Transformation and Load
- Spark for real-time analytics
- Scripted, repeatable, reliable and *fast*

Storage Best Practices

- Optimize an extensible data repository that will grow with low TCO

- Standard protocols for universal connectivity

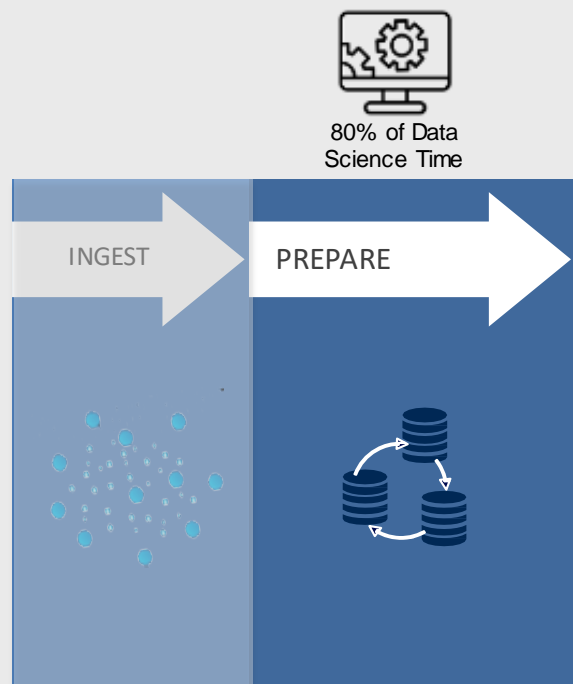- Support for structured and unstructured data

IBM Chief Data Office
4 billion client data **records** across **19** internal & external **data sources** refreshed weekly.

Serving **29,000 employees** in 60 countries @ YE18

Adopting and Expanding AI

Classify and Prepare Data Sets

80% of Data Science Time

INGEST

PREPARE

Training and Testing Data Sets
- Accuracy improves with volume of data
- Track data sets to identify bias, create related models

Make data available for other analytics
- The Data Science Toolkit includes: Hadoop, SPSS, SAS, R, etc.

Metadata Generation
- Quality improvement and regulatory compliance requires data tracking
- Use Storage and AI to automate metadata

Storage Best Practices

- Support broad analytics tools across data and metadata

- High-Speed I/O to run multiple experiments

- Tag and track data with metadata

IBM Chief Data Office
Automated Metadata Generation

90% reduction in
cycle time

Over 200k experiments used to classify TBs of data to make it discoverable

Adopting and Expanding AI

Develop AI models

Resource Optimization

INGEST → PREPARE → TRAINING →
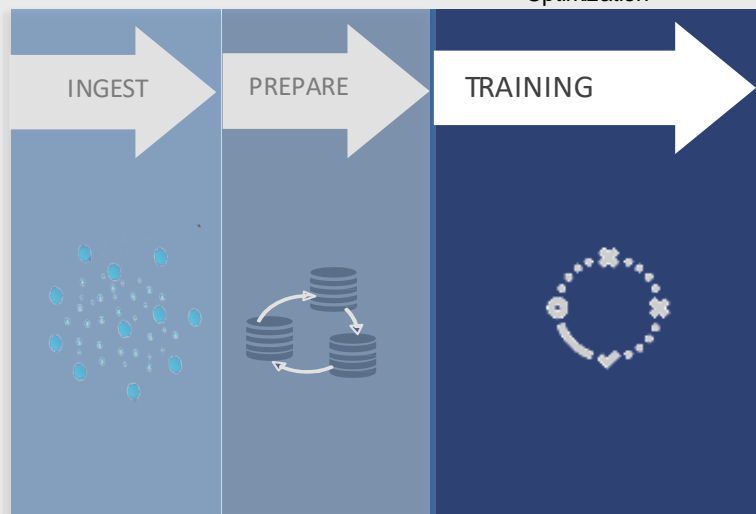
Increase Model Iterations
- Fastest copy possible to GPUs
- many models can be distributed across multiple GPUs
- Containerize models for tracking and efficiency

Leading Edge Shared Data Service
- Leading GPU can cost $10k
- GPUs : up to 150GB/s of throughput
- Shared Container Service

Storage Best Practices

- Highest throughput possible

- End-to-end Bandwidth to run multiple data models

- Local caching

- RDMA to eliminate transfer overhead

IBM Storage

Fast, Scalable Storage

2TB/s at CORAL
200GB/s in a single rack

Adopting and Expanding AI

Deploy AI models

Champion
Challenger

INGEST → PREPARE → TRAINING → INFERENCE ✓
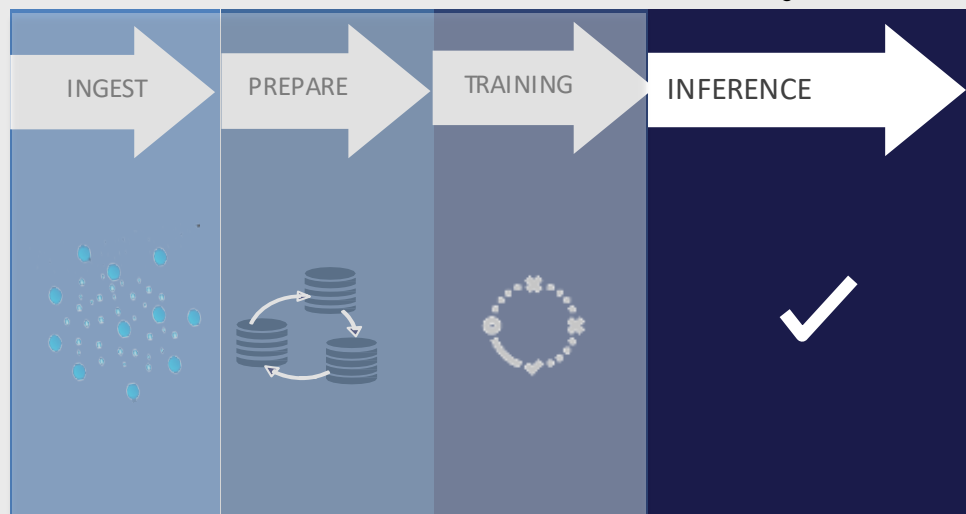
Low Latency APIs
- Fast storage for fastest response

Scalable AI service
- AI is part of a portfolio of applications
- Containerized deployments to manage workloads

Storage Best Practices

- Low latency for fastest response

- Shared API service to scale as demand is needed

- Containerized deployment

- Deploy on-premises, in cloud, or as service

IBM Chef Data Office
Quick & easy.

<1 second
response time
40,000 API calls/quarter

Presented in familiar user interface.

Building Enterprise AI
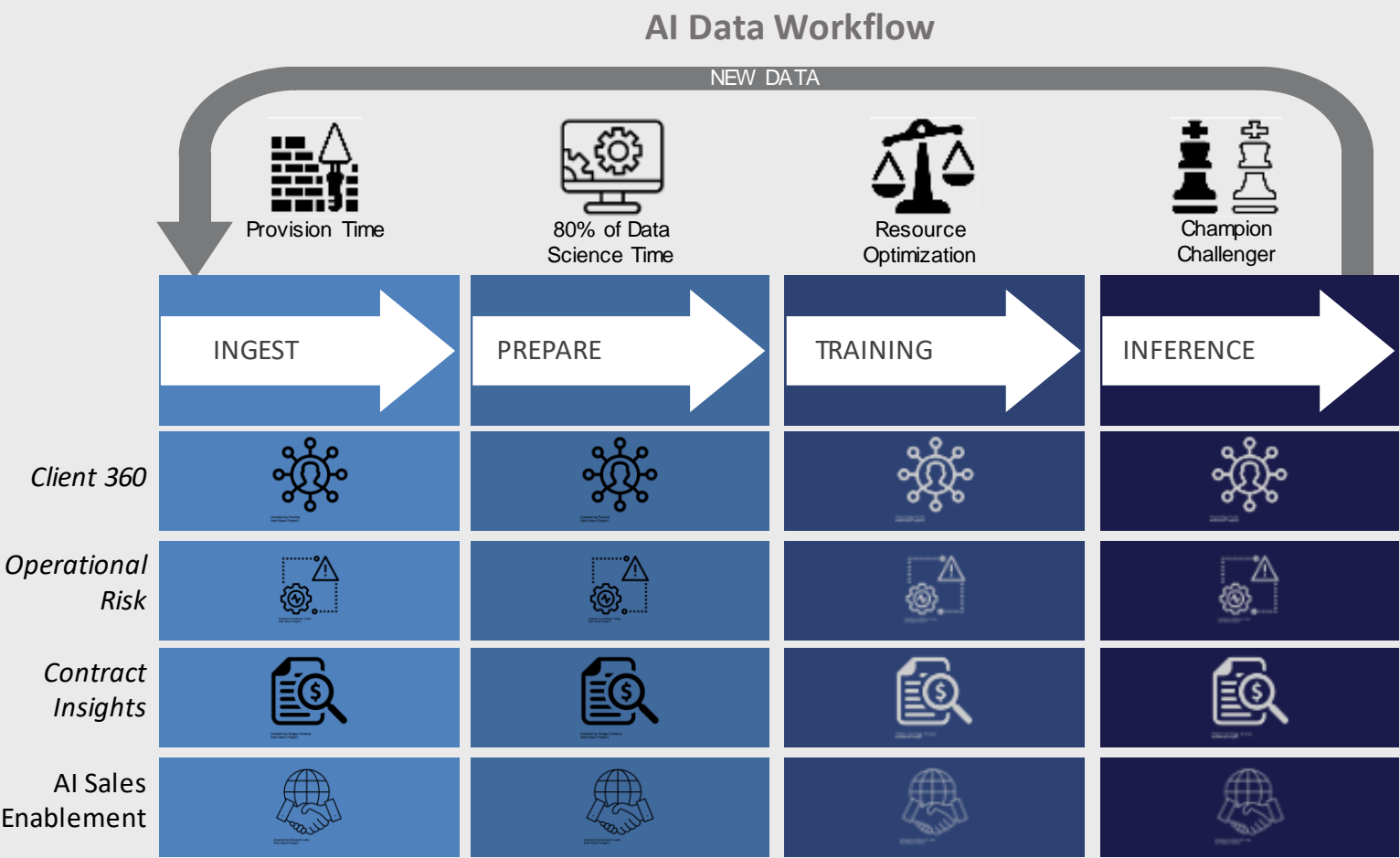
A Proof of Concept

- Easy to get started
- Can start with less data
- Demonstrate Business Value to

AI Data Workflow

NEW DATA

	Provision Time	80% of Data Science Time	Resource Optimization	Champion Challenger
	INGEST	PREPARE	TRAINING	INFERENCE
Client 360				
Operational Risk				
Contract Insights				
AI Sales Enablement				

But Think Ahead…

- Share project overhead
- Avoid Data Silos
- Address Data Governance
- Common Platforms

Enterprise AI requires Shared Data

AI Data Workflow

NEW DATA

| INGEST | PREPARE | TRAINING | INFERENCE |

UNIFIED DATA & METADATA

IBM Cloud **Object Storage**   IBM **Spectrum Discover**   IBM **Spectrum Scale**
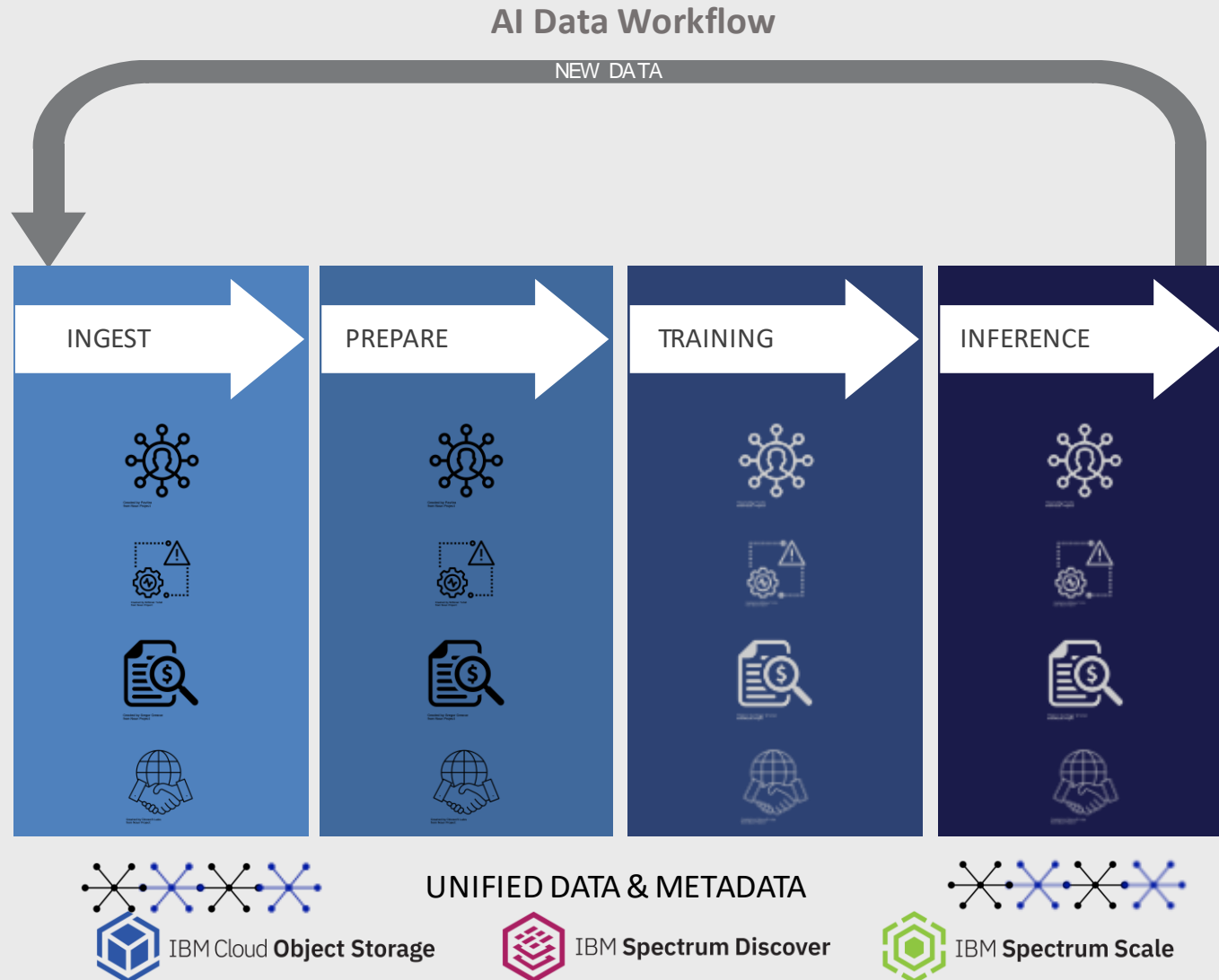
Common Enterprise Data Platform

- Single Source of Truth

- Unified Data Access

- Faster Cycle Times

- Data Network Effect accelerates AI Adoption

- Containerized, Cloud Native applications

Unique Storage Requirements
- Unified Access
  - Combine data science and analytics tools
- Performance
  - GPUs at > 100GB/s
  - Global ingest
- Smart Growth
  - Automate Tiering
  - Active Archives
- Governance
  - Extensible metadata
  - Enterprise data protection

Data Monetization:
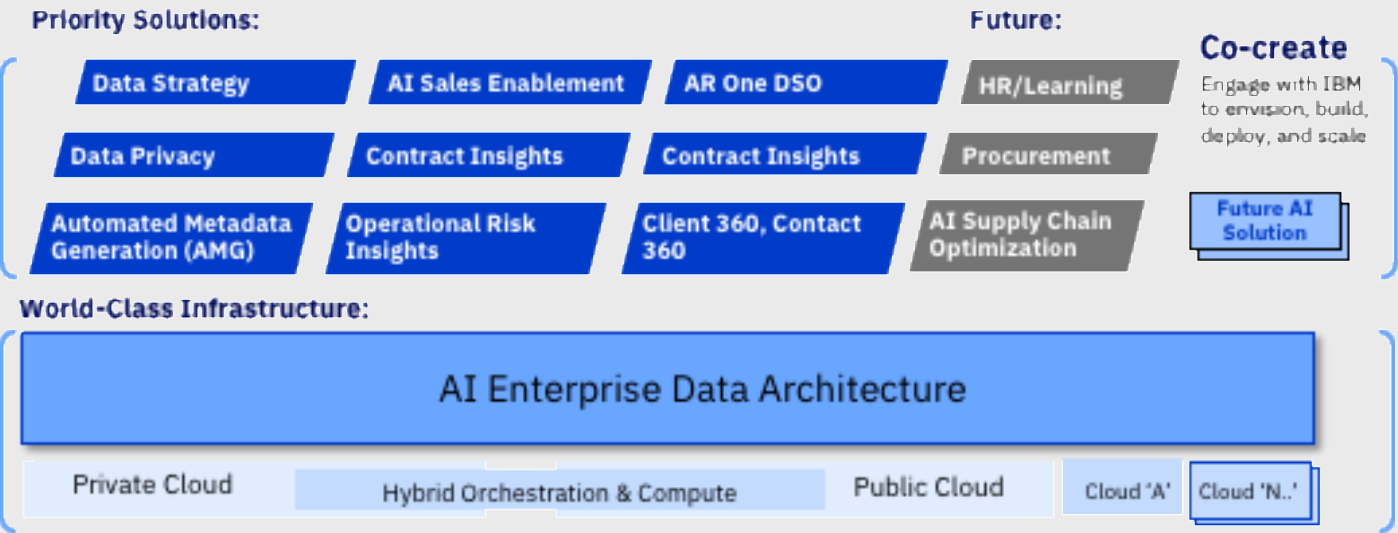# Building a Data and AI Backbone for the Enterprise

AI Accelerator

IBM is rapidly executing our AI transformation

Priority Solutions:

			Future:	**Co-create**
Data Strategy	AI Sales Enablement	AR One DSO	HR/Learning	Engage with IBM to envision, build, deploy, and scale
Data Privacy	Contract Insights	Contract Insights	Procurement	
Automated Metadata Generation (AMG)	Operational Risk Insights	Client 360, Contact 360	AI Supply Chain Optimization	Future AI Solution

World-Class Infrastructure:

AI Enterprise Data Architecture

Private Cloud	Hybrid Orchestration & Compute	Public Cloud	Cloud 'A'	Cloud 'N..'

Built upon a shared, flexible storage architecture

IBM AI Enterprise Transformation – Signature Accomplishments*

30,000 IBM users on CEDP at end of 2018

Central platform uptake **1 trillion records** (100% increase quarter over quarter)

High availability with **data queries in seconds** rather than minutes/hours

Average **60% improvement in cycle time** across IBM business processes

Dr. Steven Eliuk

IBM Global Chief Data Office

IBM Common Enterprise Data Platform

- IBM Spectrum Scale
- Elastic Storage Server (ESS)
- 1.5PB of Flash
- 4PB of disk
- Supporting analytics, AI, and scripts with IBM Cloud Private containers

IBM Spectrum Scale

* Since 4th quarter 2016 platform launch

Automated Metadata Generation

An AI-powered process for curating, verifying, and classifying data that enhances speed and usability at speed.

Co-creation partner:
IBM Analytics

Project Contact:
Sonia Mezzetta

Challenge:
An overwhelming amount of data can leave important information in the dark, inhibiting insights and ultimately business value.

Sensitive data is often either blocked or at risk.

Classifying data for discoverability is a labor intensive manual process that can take weeks.

Unified.
Classifying **terabytes of data** to make it **easily discoverable**

While providing the **data stewardship, lineage, and impact analysis** to assure it is **trustworthy.**
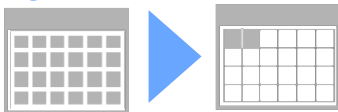
Quick & easy.
Automated data curation in days instead of weeks.

Backed by micro-services to **develop once, run anywhere**

Smart.
Using **Watson** for NLC

& custom **Deep Learning** for metadata classifications

Up to **90% reduction in cycle time:**

$27 million in **productivity savings**

Dramatically enhanced **Data Quality** with regulatory & governance checks

Supply Chain Risk Insights

An automated, comprehensive and Watson-powered Alert Service that assesses supply chain risk events to identify those posing the greatest threat of impact

Partners:
CAO
Procurement
GTS Resiliency Services
Systems Supply Chain
RESO
IGF

Project Contact:
Tom Ward

Challenge:
Global supply chains subject to high costs from missing a serious threat or overreacting to a minor one.

Multiple separate Risk Alert data feeds, warnings, watches and advisories are impossible for an individual to synthesize and assess.

Unified.
150 data sources
monitored hourly to synthesize **millions of records in real-time**

Analyze risk-event impact to
12,500+ supply chain and data center sites
worldwide

Quick & easy.
Single, constantly-updating view of world events and IBM locations to **instantly notify affected sites**

$9 million gain
in efficiency & cost avoidance in two years

The Weather Company
An IBM Business

#1 internal user

Smart.
Watson tells us **when to respond or ignore**

Client 360

A single enterprise standard and certified view of client data, enhanced with AI-powered insights

Challenge:
Silos of IBM client information with different identifiers

Excessive time to manually collect client interactions

Incomplete client picture

No way to enable AI-powered insights

Unified.

4 billion client data **records**

across **12***

internal & external

data sources

refreshed weekly.

Serving **29,000 sellers** in 60 countries by YE18

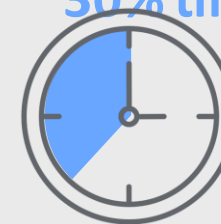**Quick & easy.**

<1 second
response time
40,000 API calls/quarter
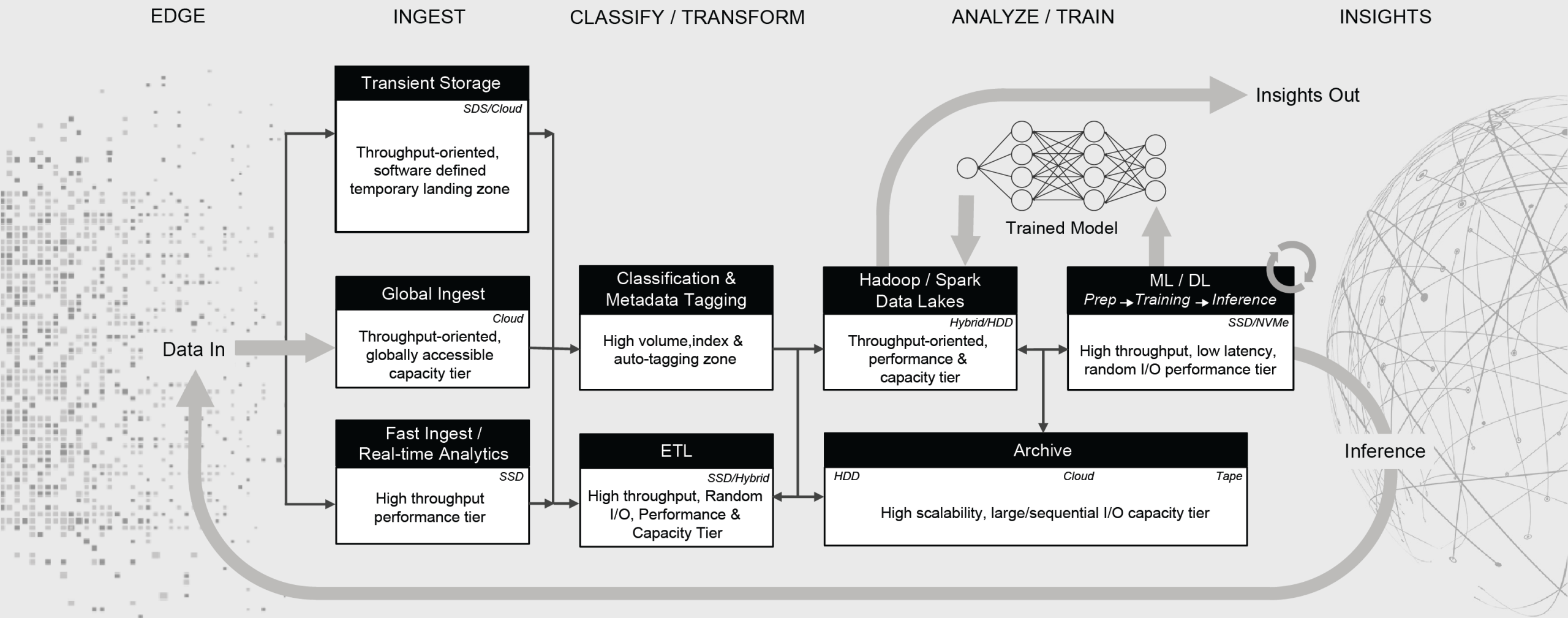
Presented in familiar user interface.

30% time savings
for sellers

Smart.
Actionable recommendations by territory and industry
Coming soon!

* 19 data sources by 3Q18

Data Pipeline with Storage Requirements

EDGE · INGEST · CLASSIFY / TRANSFORM · ANALYZE / TRAIN · INSIGHTS

Transient Storage
SDS/Cloud

Throughput-oriented, software defined temporary landing zone

Global Ingest
Cloud

Throughput-oriented, globally accessible capacity tier

Fast Ingest / Real-time Analytics
SSD

High throughput performance tier

Classification & Metadata Tagging

High volume, index & auto-tagging zone

ETL
SSD/Hybrid

High throughput, Random I/O, Performance & Capacity Tier

Hadoop / Spark Data Lakes
Hybrid/HDD

Throughput-oriented, performance & capacity tier

Archive
HDD · *Cloud* · *Tape*

High scalability, large/sequential I/O capacity tier

ML / DL
Prep → Training → Inference
SSD/NVMe

High throughput, low latency, random I/O performance tier

Data In

Insights Out

Trained Model

Inference

Solutions – IBM Spectrum Storage for AI

IBM Spectrum Storage for AI supercharges your AI data pipeline with **storage solutions optimized for the unique demands of AI**. Integrating industry-leading servers, ISV / open source software and IBM software-defined storage, IBM Spectrum Storage for AI delivers simplified deployment, groundbreaking performance, and extended data management to drive developer productivity with the fastest path to insights.

IBM Spectrum Storage for AI – Available Solutions

https://www.ibm.com/it-infrastructure/storage/ai-infrastructure

- IBM Spectrum Storage for Hadoop/Spark workloads
    - IBM Spectrum Scale and Hortonworks/Cloudera Integration
    - IBM Spectrum Scale and IBM Spectrum Conductor for Spark Integration

- IBM Spectrum Storage for AI with NVIDIA DGX
    - IBM Spectrum Scale and NVIDIA DGX Reference Architecture

- IBM Spectrum Storage for AI with Power Systems
    - IBM Spectrum Scale and Power AC922 Reference Architecture

- IBM Spectrum Connect – Storage Enabler for Containers

- IBM Spectrum Storage for AI in Autonomous Driving