Accelerating AI workloads with IBM Spectrum Scale

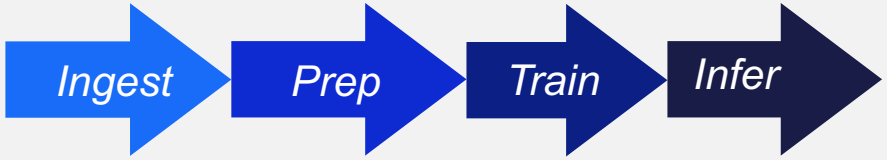
—

Ted Hoover
Program Director
IBM Spectrum Scale Development

IBM

AI “Industrial Revolution”

Common & Standard Workflow Using AI Innovations

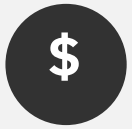


AI Examples in Every Industry

Autonomous driving Accident avoidance

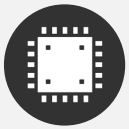
Location-based advertising

Sentiment analysis of what's hot, problems



Market prediction Fraud/Risk

Experiment sensor analysis



Mfg. quality Warranty analysis

Clinical trials, drug discovery, Genomics

Captioning, search, real time translation

People & career matching

Patient sensors, medical image interpretation

Drilling exploration sensor analysis

Consumer sentiment Analysis

Sensor analysis for optimal traffic flows

Smart Meter analysis for network capacity,

Threat analysis - social media monitoring, video Surveillance

Better Storage Removes the Impediments to AI Insights

Driven by Data

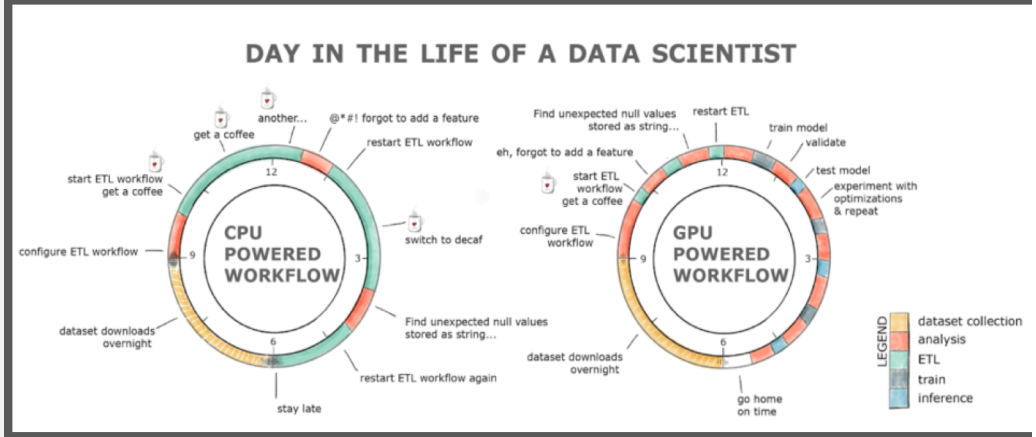
- The fidelity of Machine Learning and Deep Learning is absolutely dependent having access to larger quantities of high quality data.

Readily Accessible

- Majority of the data scientists' time is spent on data collection, transform and selection

Always Growing

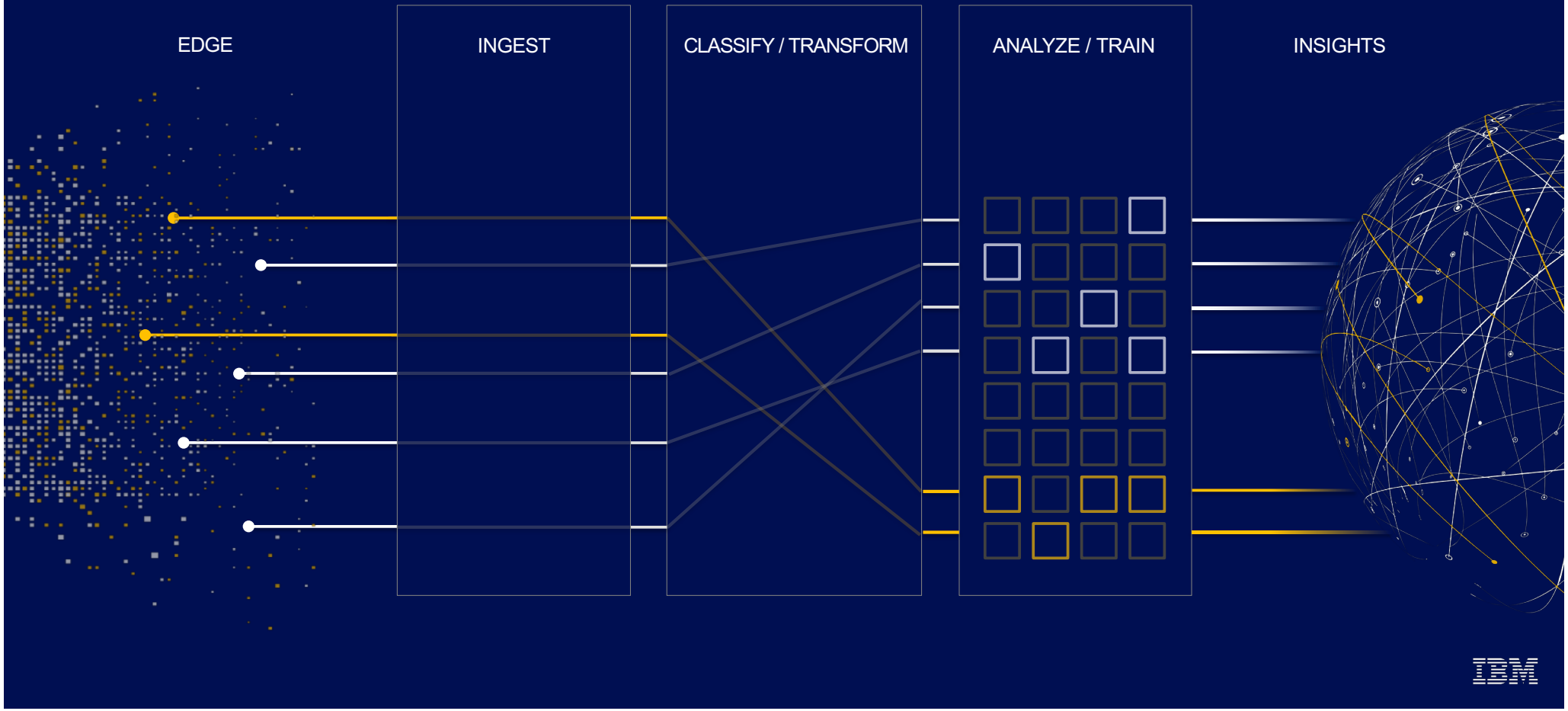
- Improving AI models is iterative and dependent upon a growing set of data for training and testing



https://devblogs.nvidia.com/gpu-accelerated-analytics-rapids/

The Goal: *Move Data from Ingest to Insights*

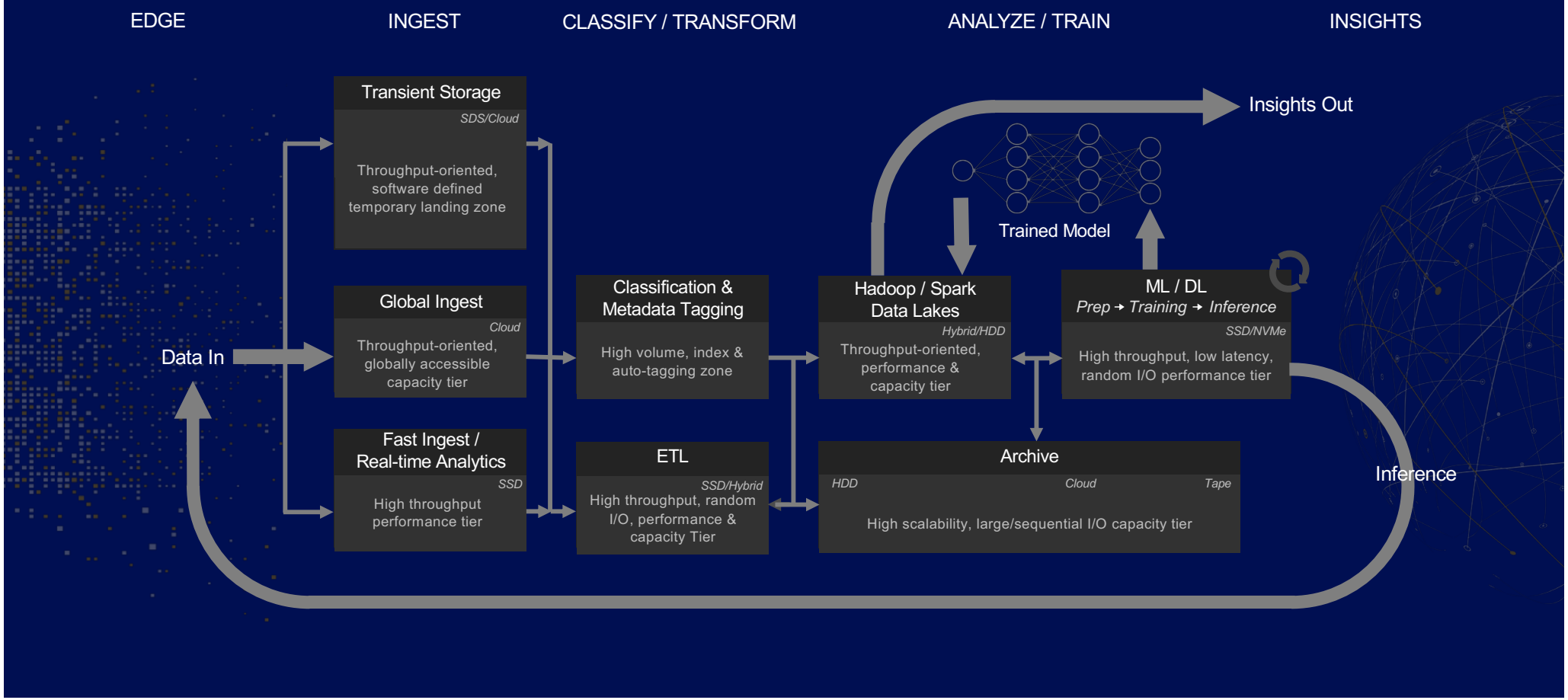
IBM Storage and SDI



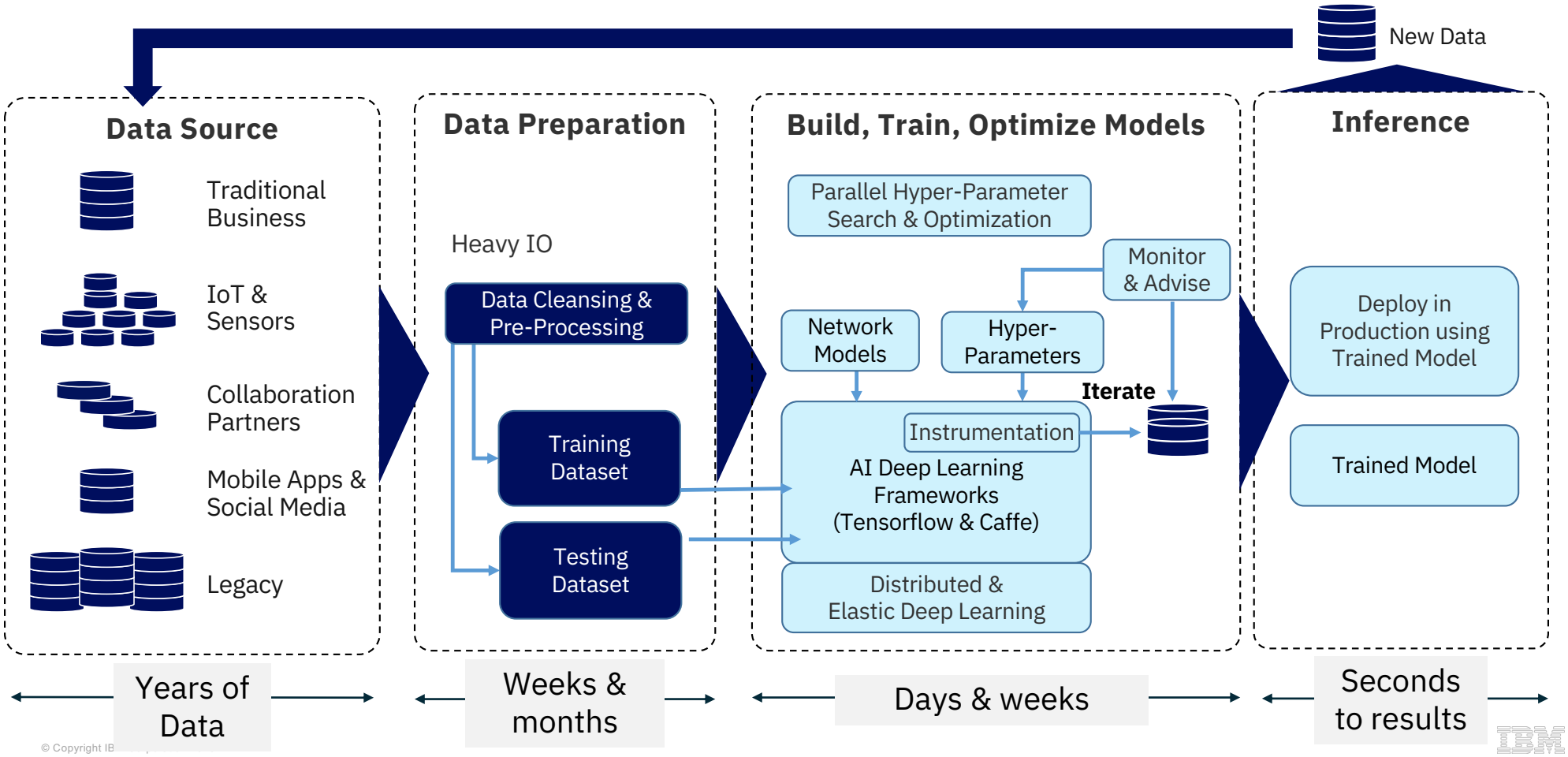
IBM

AI Data Pipeline

IBM Storage and SDI



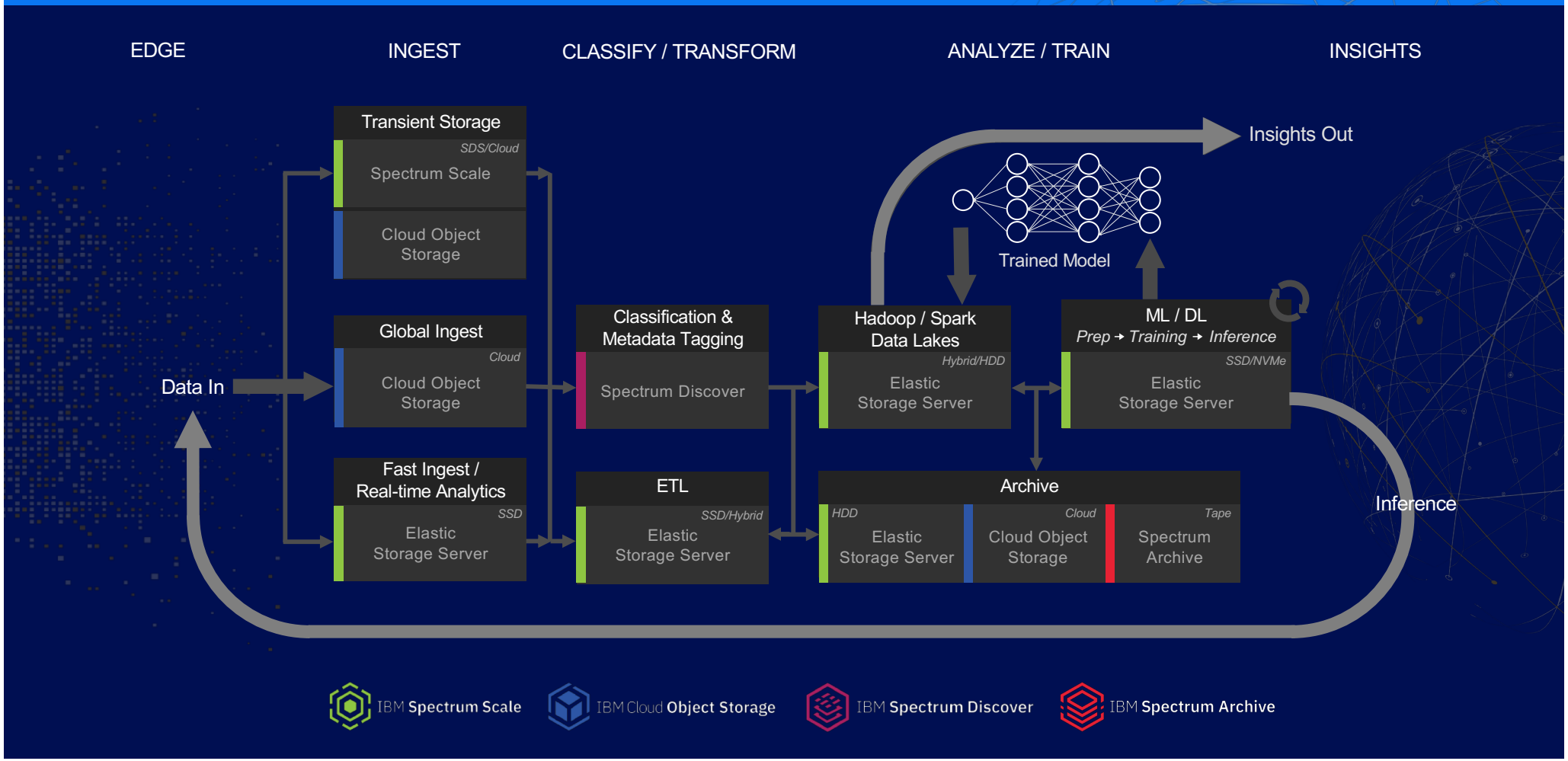
Workflow and Data Flow is Complex



AI Data Pipeline with Spectrum Storage

Improved data governance with storage offerings for end-to-end data pipeline

IBM Storage and SDI



Storage Requirements

Ingest and Preparation

- **Requirement #1: Extremely High Throughput**

Model Training

- Requirement #2: Extremely High Throughput & Low Latency

Model Inference

- Requirement #3: Extremely Low Latency

Active and Cold Archive

- Requirement #4 Extremely Large Scale

Capacity and Performance tier: Data Collection, Data aggregation and Normalization

- Throughput-oriented
- Protocols: SMB, NFS, HDFS
- Edge, On-Prem, Cloud

Spectrum Scale Features / Improvements:

- Muti-protocol support
- 2.5 TB/sec - Throughput to storage architecture (CORAL)
- Hybrid Cloud

Storage Requirements

Ingest and Preparation

- Requirement #1: Extremely High Throughput

Model Training

- **Requirement #2: Extremely High Throughput & Low Latency**

Model Inference

- Requirement #3: Extremely Low Latency

Active and Cold Archive

- Requirement #4 Extremely Large Scale

Training: Performance tier: Model training by parallelization of processes

- High bandwidth, low latency, small random I/O
- Protocols: SMB, NFS
- On-Prem, Public Cloud

Spectrum Scale Features / Improvements:

- Mmap() performance enhancements
- Further IO optimization for write-once-but-read-many-times
- Prefetch/cache all data in LROC locally (in progress)

Storage Requirements

Ingest and Preparation

- Requirement #1: Extremely High Throughput

Model Training

- Requirement #2: Extremely High Throughput & Low Latency

Model Inference

- **Requirement #3: Extremely Low Latency**

Active and Cold Archive

- Requirement #4 Extremely Large Scale

Inference: Performance tier: Model analyses

- Low Latency, mixed read/write Workloads
- Protocols: SMB, NFS
- On-Prem, Public Cloud

Spectrum Scale Features / Improvements:

- Tool Integration – Data Pipeline
- Best Practices Guide (in progress)
- Tuning profile for cognitive workloads

Storage Requirements

Ingest and Preparation

- Requirement #1: Extremely High Throughput

Model Training

- Requirement #2: Extremely High Throughput & Low Latency

Model Inference

- Requirement #3: Extremely Low Latency

Active and Cold Archive

- **Requirement #4 Extremely Large Scale**

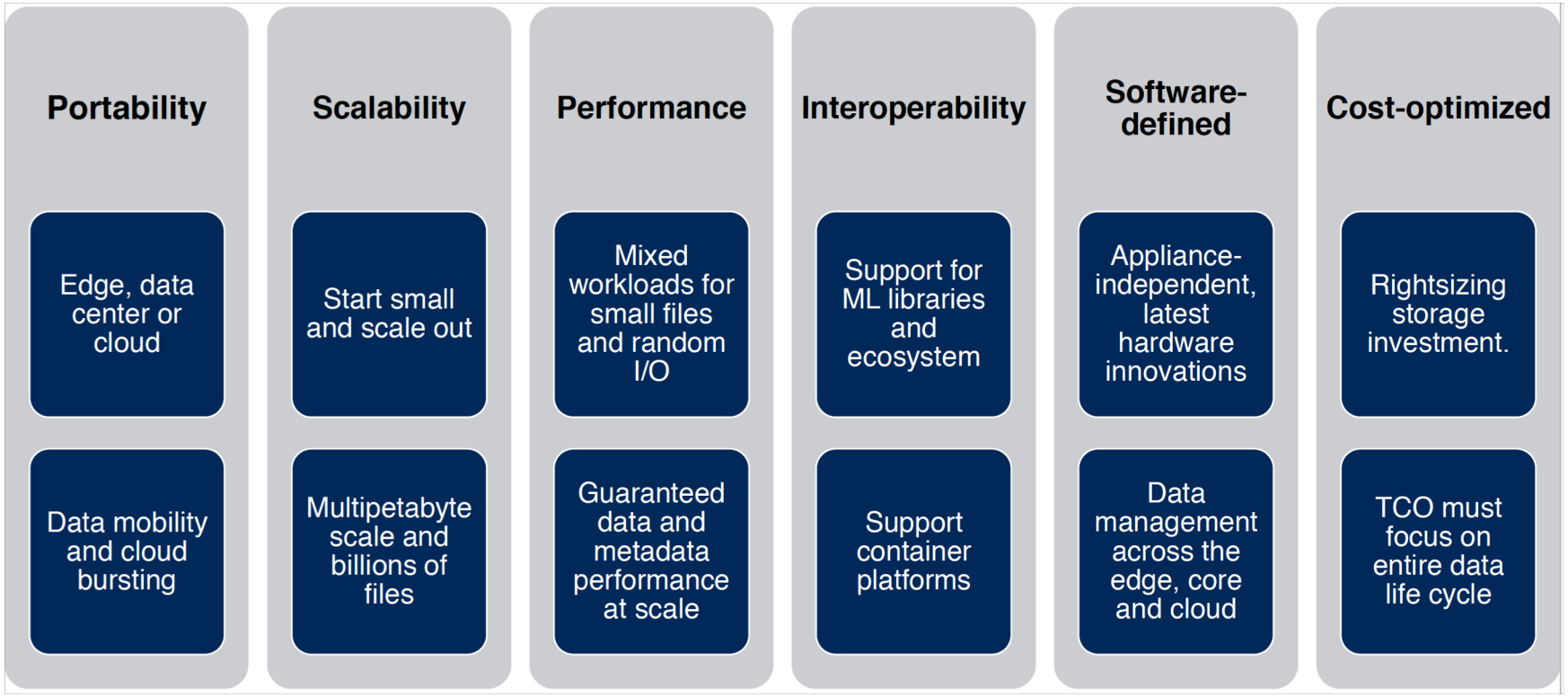
Archive: Capacity Tier: active and cold archive

- Throughput-oriented, large I/O, streaming, sequential writes
- Protocols: S3,LTFS
- On-Prem, Public Cloud

Spectrum Scale Features / Improvements:

- Data Tiering - ILM
- Spectrum Archive Integration
- Data Tiering to Cloud

Storage Selection Requirements



ibm.com/storage

Thank you