Accelerating AI workloads with IBM Spectrum Scale

Ted Hoover Program Director IBM Spectrum Scale Development



AI "Industrial Revolution"











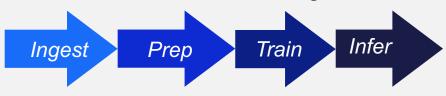








Common & Standard Workflow Using Al Innovations







Predictive Analytics





Physics



Customer Experience



Cybersecurity Fraud Detection



AI Examples in Every Industry



Autonomous driving Accident avoidance



Mfg. quality Warranty analysis



Drilling exploration sensor analysis



Location-based advertising



Sentiment analysis of what's hot, problems



Market prediction Fraud/Risk



Experiment sensor analysis



Clinical trials, drug discovery, Genomics



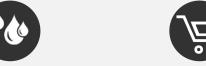
Captioning, search, real time translation



People & career matching



Patient sensors, medical image interpretation



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.

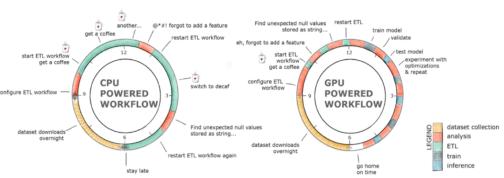
Readily Accessible

 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

DAY IN THE LIFE OF A DATA SCIENTIST

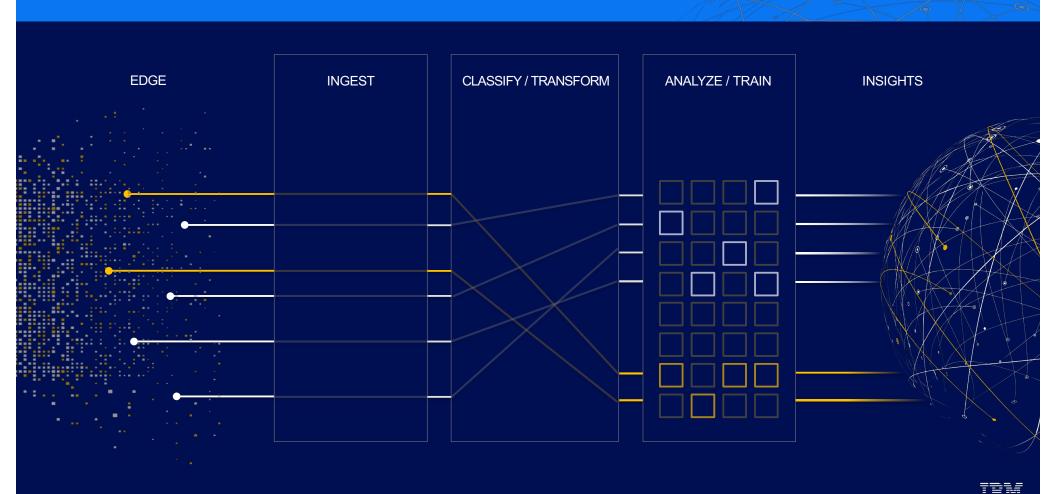


https://devblogs.nvidia.com/gpu-accelerated-analytics-rapids/

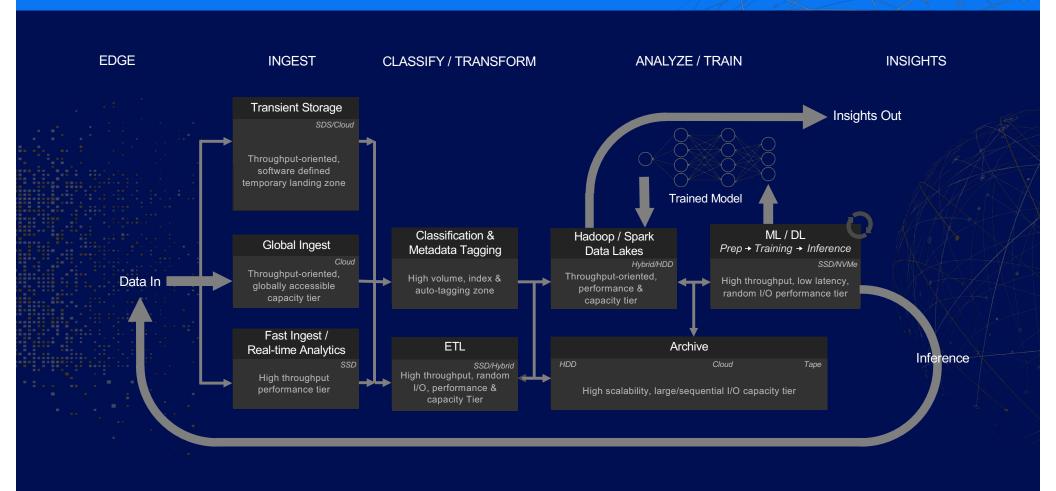
© Copyright IBM Corporation 2018

IBM Storage and SDI

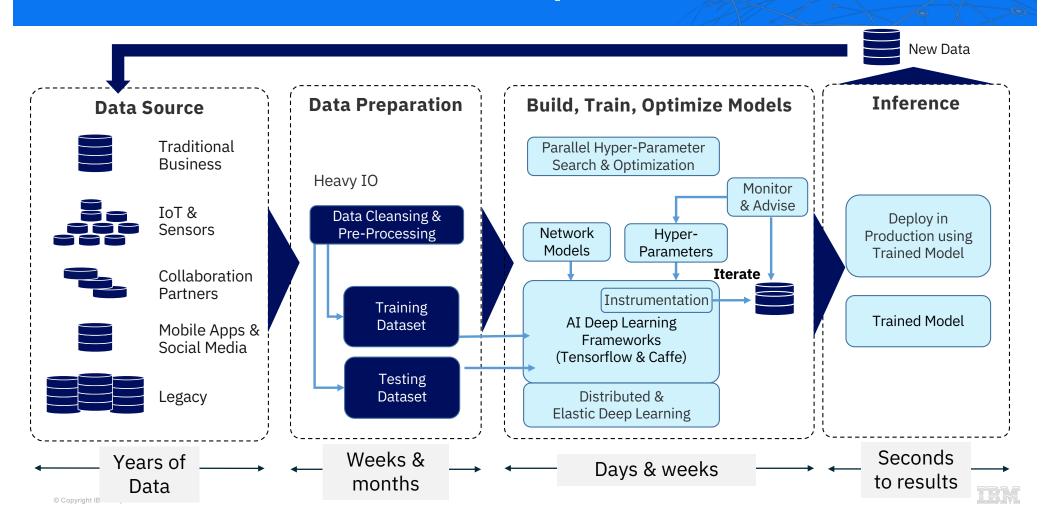
The Goal: Move Data from Ingest to Insights



AI Data Pipeline

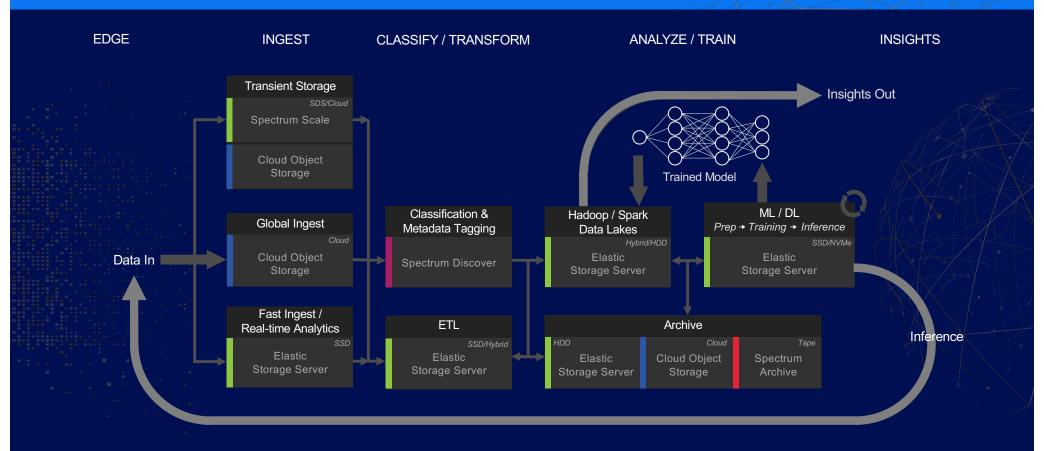


Workflow and Data Flow is Complex



AI Data Pipeline with Spectrum Storage

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











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

© Conveight IRM Corporation 20

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-readmany-times
- Prefetch/cache all data in LROC locally (in progress)

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

© Copyright IBM Corporation 201

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

© Copyright IBM Corporation 201

Storage Selection Requirements

Portability

Edge, data center or cloud

Data mobility and cloud bursting

Scalability

Start small and scale out

Multipetabyte scale and billions of files

Performance

Mixed workloads for small files and random I/O

Guaranteed data and metadata performance at scale

Interoperability

Support for ML libraries and ecosystem

Support container platforms

Softwaredefined

Applianceindependent, latest hardware innovations

Data management across the edge, core and cloud

Cost-optimized

Rightsizing storage investment.

TCO must focus on entire data life cycle

© 2018 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

ibm.com/storage

Thank you

