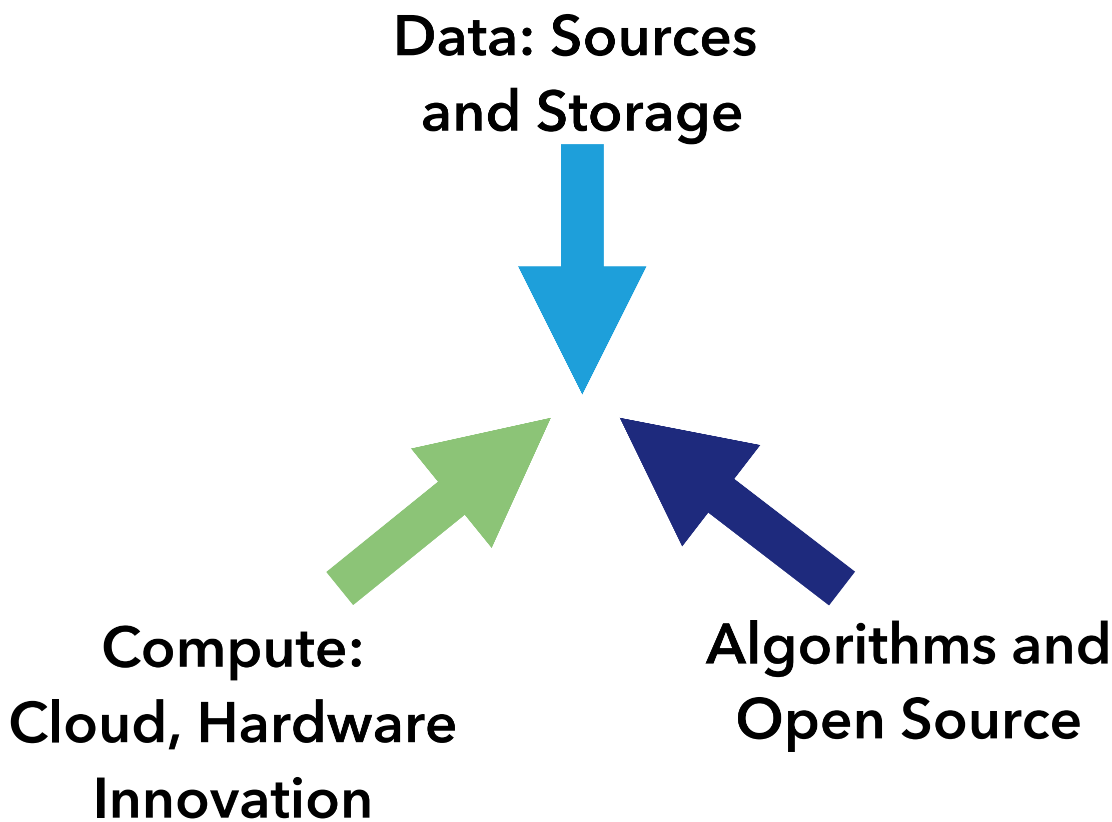
Machine Learning and Storage Applications

Nisha Talagala

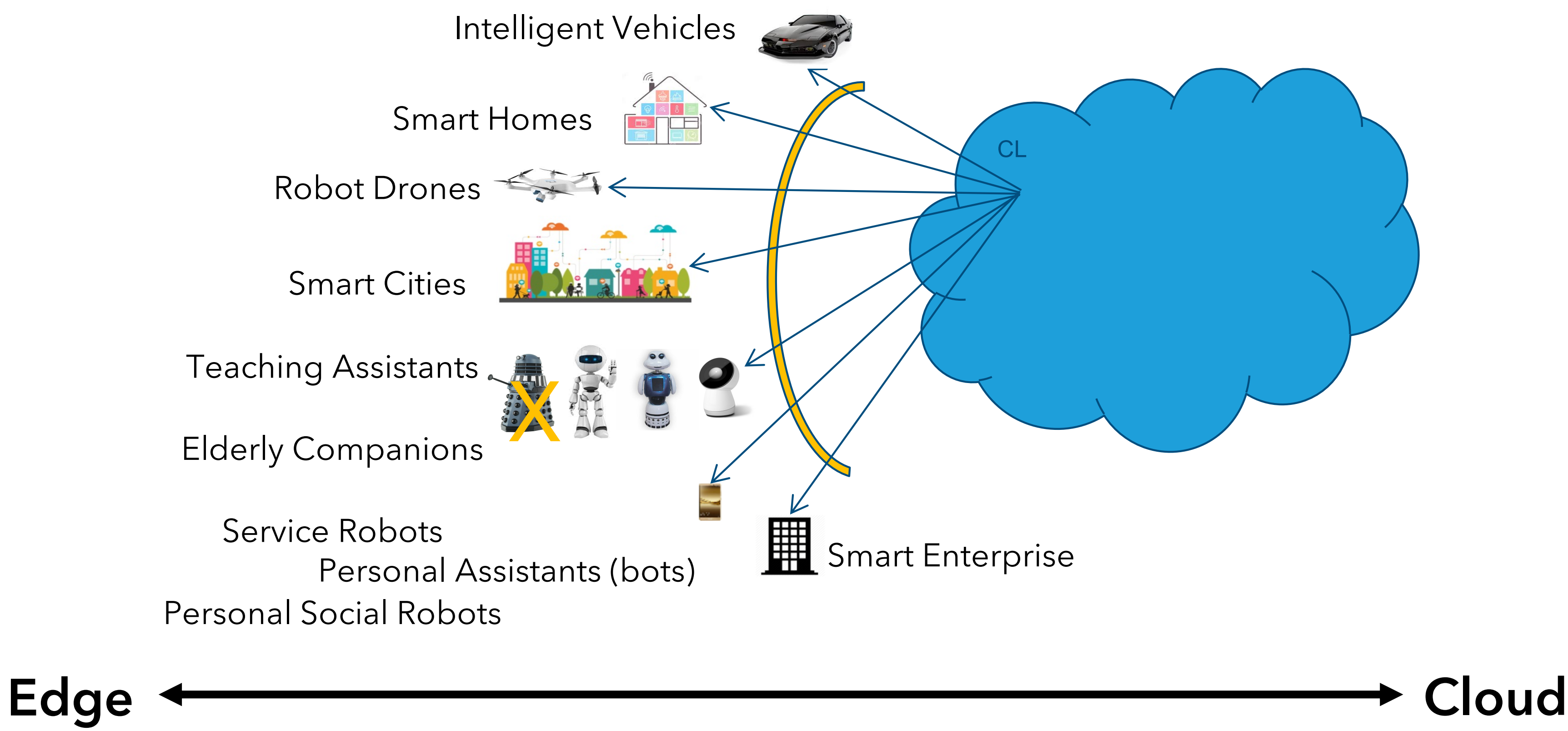
CTO, ParallelM

Flash Memory Summit 2018

Machine Learning Growth

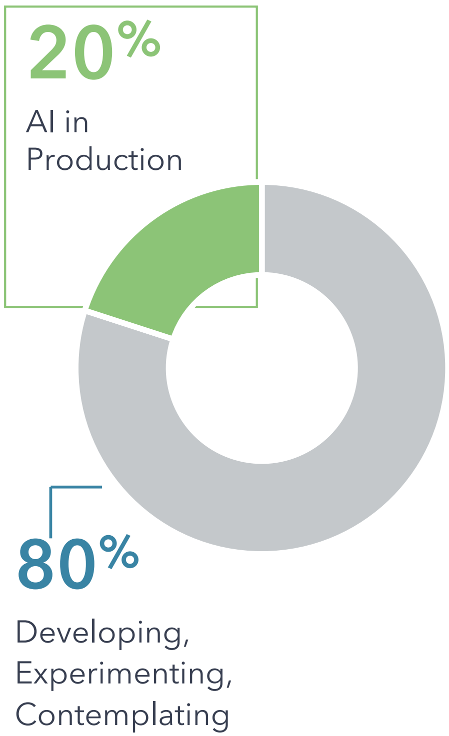


Growing Sources of Data



Edited version of slide from Balint Fleischer's talk:
Flash Memory Summit 2016, Santa Clara, CA

Growing AI Investments but Few Deployed at Scale



Survey of 3073 AI-aware C-level Executives

Out of 160 reviewed AI use cases:

88% did not progress beyond the experimental stage

But successful early AI adopters report:

Profit margins **3–15%** higher than industry average

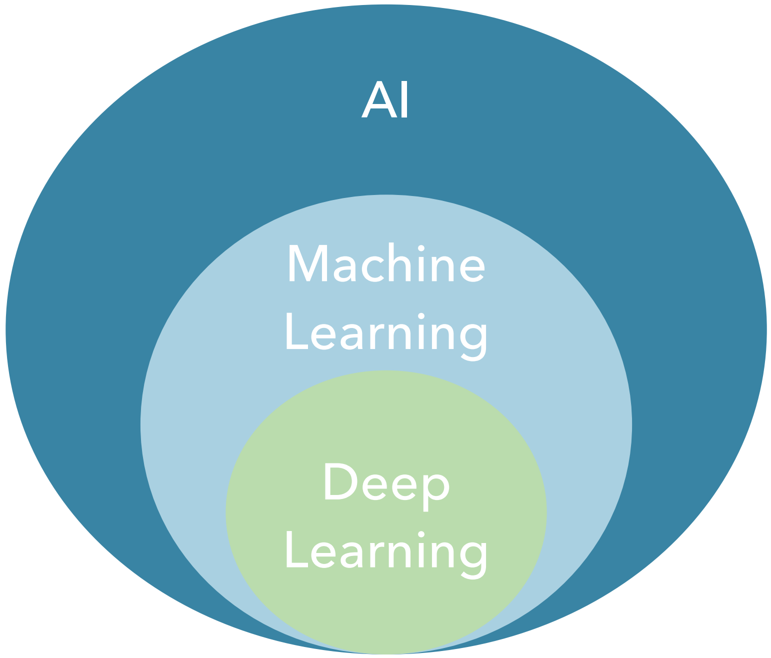
Source: "Artificial Intelligence: The Next Digital Frontier?", McKinsey Global Institute, June 2017

In This Talk:

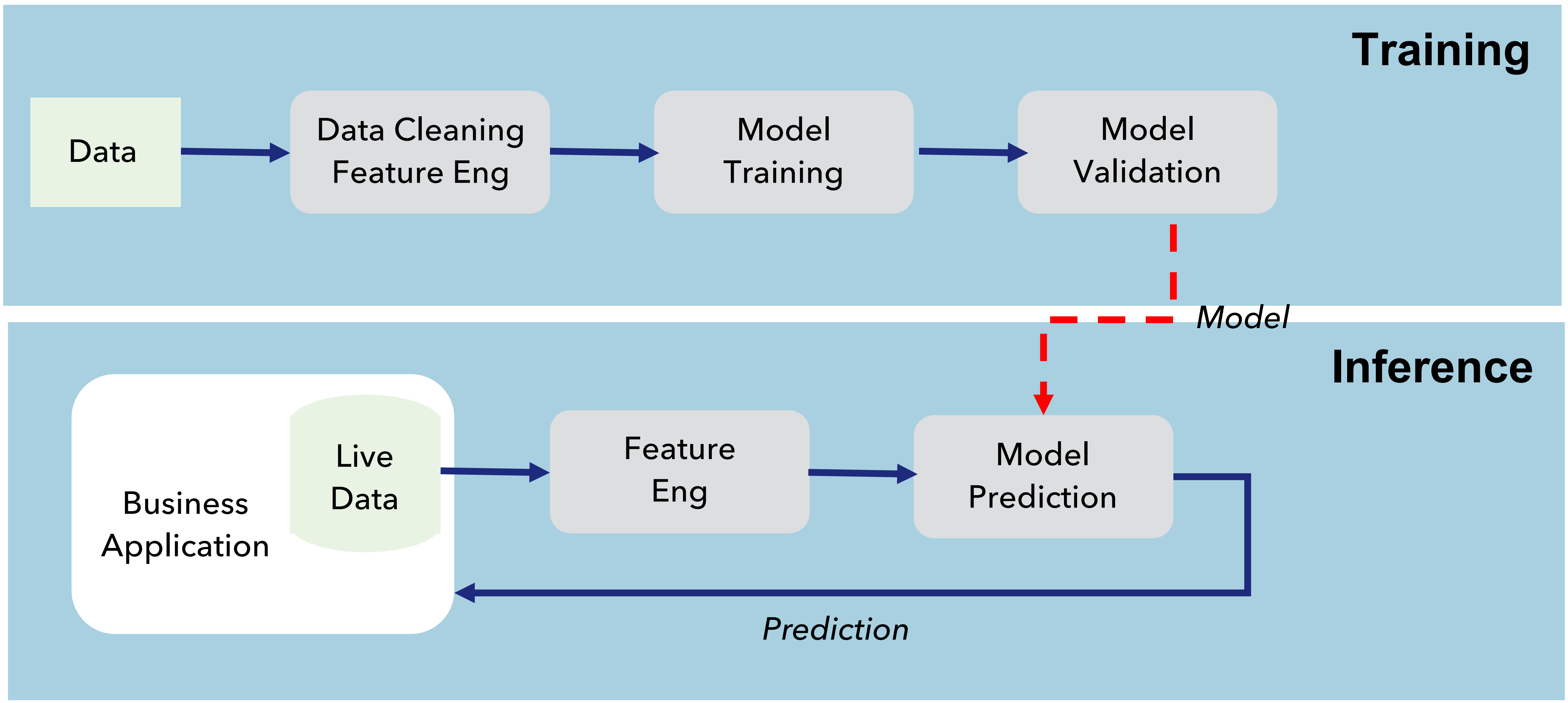
- AI and ML: A quick overview
- ML Application trends as relevant for Storage
- Opportunities for using ML inside Storage

What is Machine Learning and AI?

- AI: Natural Language Processing, Image Recognition, Anomaly Detection, etc.
- Machine Learning: Supervised, Unsupervised, Reinforcement, Transfer, etc.
- Deep Learning: CNNs, RNNs etc.
- Common Threads
  - Training
  - Inference (aka Scoring, Model Serving, Prediction)



A Typical ML Operational Pipeline



A Sample Analytics Stack

Libraries	Machine Learning, Deep Learning, SQL, Graph, CEP, etc.	Libraries

Processing Engine

Data

Data Streams

FLUME

RabbitMQ

kafka

nifi

Data

Data Repositories

NoSQL

SQL

hadoop

Spark

S3

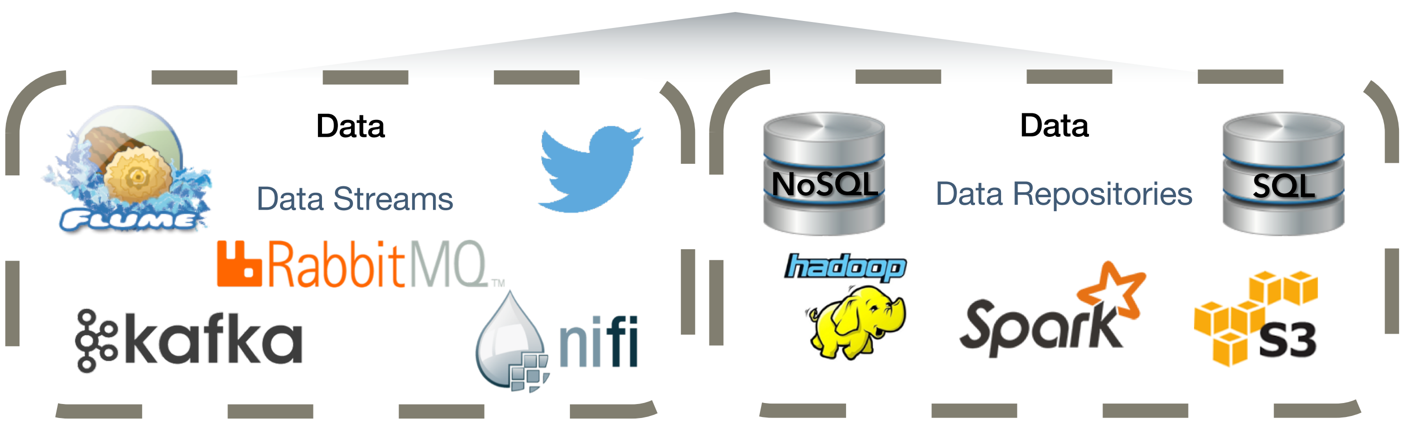
Data from Repositories or Live Streams

Parallel/\

A Sample Analytics Stack: (Partial) Ecosystem

Algorithms and Libraries	SparkML, TensorFlow

Processing Engines	Hadoop Spark Tensor Flow	Flink / Apex Spark Streaming Storm / Samza / NiFi	Caffe Tensor Flow Pytorch	Containerized Models (Python etc.)



Data from Repositories or Live Streams

Trend 1: How ML/DL Workloads Think About Data

- Data Sizes
  - Incoming datasets can range from MB to TB
  - Models are typically small. Largest models tend to be in deep neural networks and range from 10s MB to single digit GB
- Common Structured Data Types
  - Time series and Streams
  - Multi-dimensional Arrays, Matrices and Vectors
- Common distributed patterns
  - Data Parallel, periodic synchronization
  - Model Parallel
  - Straggler performance issues can be significant

Trend 1: How ML/DL Workloads Think About Data

- The older data gets – the more its “role” changes
  - Older data for batch- historical analytics and model reboots
  - Used for model training (sort of), not for inference
- Guarantees can be “flexible” on older data
  - Availability can be reduced (most algorithms can deal with some data loss)
  - A few data corruptions don’t really hurt ☺
  - Data is evaluated in aggregate and algorithms are tolerant of outliers
  - Holes are a fact of real life data – algorithms deal with it
- Quality of service exists but is different
  - Random access is very rare
  - Heavily patterned access (most operations are some form of array/matrix)
  - Shuffle phase in some analytic engines

Trend 2: Need for Governance

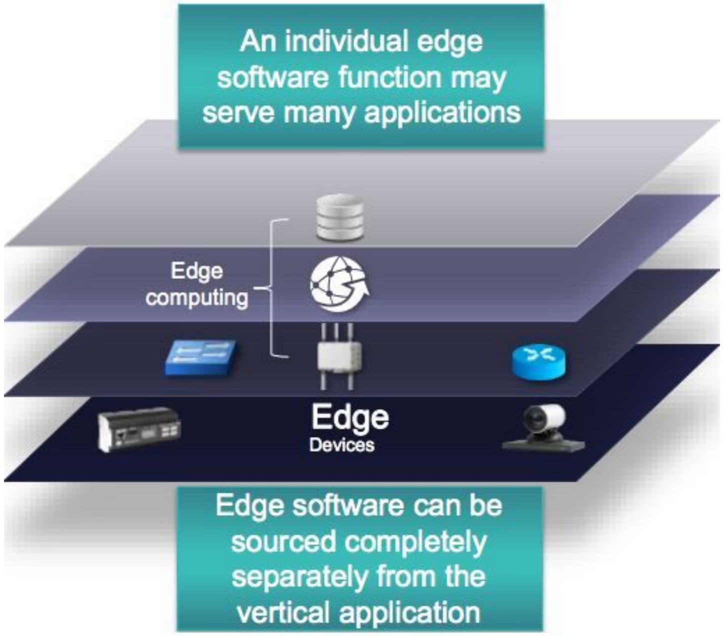
- Examples
  - Established: Example: Model Risk Management in Financial Services
  - https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf
- Emerging: Example GDPR on Reproducing and Explaining ML Decisions
  - https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/
- Emerging: New York City Algorithm Fairness Monitoring
  - https://techcrunch.com/2017/12/12/new-york-city-moves-to-establish-algorithm-monitoring-task-force/

Trend 2: Need for Governance

- ML is only as good as its data
- Managing ML requires understanding ***data provenance***
  - *How was it created? Where did it come from? When was it valid?*
  - *Who can access it? (all or subsets)? Which features were used for what?*
  - *How was it transformed?*
  - *What ML was it used for and when?*
- Solutions require both storage management and ML management

Trend 3: The Growing Role of the Edge

- Closest to data ingest, lowest latency.
  - Benefits to real time ML inference and (maybe later) training
- Varied hardware architectures and resource constraints
- Differs from geographically distributed data center architecture
- Creates need for cross cloud/edge data storage and management strategies



IoT Reference Model

Trend 4: The Changing Role of Persistence

- For ML functions, most computations today are in-memory
  - Data load and store are primary storage interaction
  - Intermediate data storage sometimes used
  - Tiered memory can be used within engines
- For in-memory databases, persistence is part of the core engine
  - Log based persistence is common
- Loading & cleaning of data is still a very large fraction of the pipeline time
  - Most of this involves manipulating stored data

Trend 5: The Growth of Streaming Data

- Continuous data flows and continuous processing
- Enabled & driven by sensor data, real time information feeds
- Several variants with varied functionality
  - True Streams, Micro-Batch (an incremental batch emulation)
- The performance of in-memory streaming enables a convergence between stream analytics (aggregation) and Complex Event Processing (CEP)
- Bring need for stream optimized data stores

Trend 6: The ML Job Functions

- Multiple ML roles interact with data
  - Data Scientist
  - Decision Scientist
  - Data Engineer
- ML roles need to collaborate with Operations roles for successful Operational ML.
- Requires data access controls, access management to ensure ML consistency and governance

Storage for ML: Challenges and Opportunities

- Ingest Speeds (Particularly for Deep Learning Workloads)
- Data Management for ML Workloads
- Governance and the Challenges of Regulation, Data Access Control and Access Management
- The Edge
- Streaming Data

Storage for ML: Examples

- RDMA data acceleration for Deep Learning (Ex. from Mellanox)
- Time series optimized databases (Ex. BTrDB, GorrillaDB)
- API pushdown techniques and Native RDD Access APIs (Ex. Iguaz.io)
- Lineage: Link data and compute history (Ex. Alluxio/formerly Tachyon)
- Memory expansion (Ex. Many studies on DRAM/Persistent Memory/Flash tiering for analytics)

In This Talk:

- AI and ML: A quick overview
- ML Application trends as relevant for Storage
- Opportunities for using ML inside Storage

ML for Storage: How to Use ML to Improve Storage?

- Caching
  - Adapting caching policy using online learning can have significant benefits
- Workload classification
  - Quantify similarity between workloads
  - Track workload changes
  - Learning workload mixes
- Learning for storage tuning
  - Data distribution / tiering
  - Reconfiguration of parameters, tiers, placement and layout
- Failure Prediction

**Taken from NSF Vision Workshop AI and Storage subteam report*

ML for Storage: How to Use ML to Improve Storage?

- Challenges
  - Training data may be limited before decisions must be made
  - Historical data is helpful. Telemetry data can be used for this purpose but telemetry may need to be adjusted
  - Firmware revs/data format changes
  - Production ML deployment and tuning is a challenge

- Examples of ML use in Storage
  - Public industry usages from Pure Storage, Netapp, etc.
  - Research examples (algorithms, experiments) published at HotStorage Workshops, FAST (File and Storage Technologies) Conferences

Takeaways

- The use of ML/DL in enterprise is at its infancy
  - Some requirements understood, many still emerging
- These apps put ever larger pressure on performance, data management and provenance
- Opportunities exist to significantly improve storage and memory for these use cases by understanding and exploiting their priorities and non-priorities for ML data

Additional Resources

- Operational ML at http://www.mlops.org
  - Articles on the practical challenges of ML in production, including provenance and governance
- Netapp: Lessons Learned Processing 70 Billion Datapoints in the Hybrid Cloud (https://www.slideshare.net/Hadoop_Summit/lessons-learned-processing-70-billion-data-points-a-day-using-the-hybrid-cloud)
- Mellanox/ParallelM – Deep Learning acceleration/management (http://www.mellanox.com/blog/2017/10/teslas-autopilot-teaches-us-devops-high-performance-ai-powered-applications/)
- Upcoming NSF Vision report on Storage for 2025
- Research at HotStorage, FAST, USENIX ATC

ParallelM

Thank You

nisha.talagala@parallelm.com