The Impact of Artificial Intelligence on Storage and IT

A SNIA EMEA Webcast

Today's Presenters

Paul Talbut
General Manager, SNIA EMEA

Glyn Bowden
Chief Architect, AI & Data Science Practice
HPE

Alex McDonald
Chair, SNIA EMEA
NetApp

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

SNIA-At-A-Glance

185
industry leading organizations

2,000
active contributing members

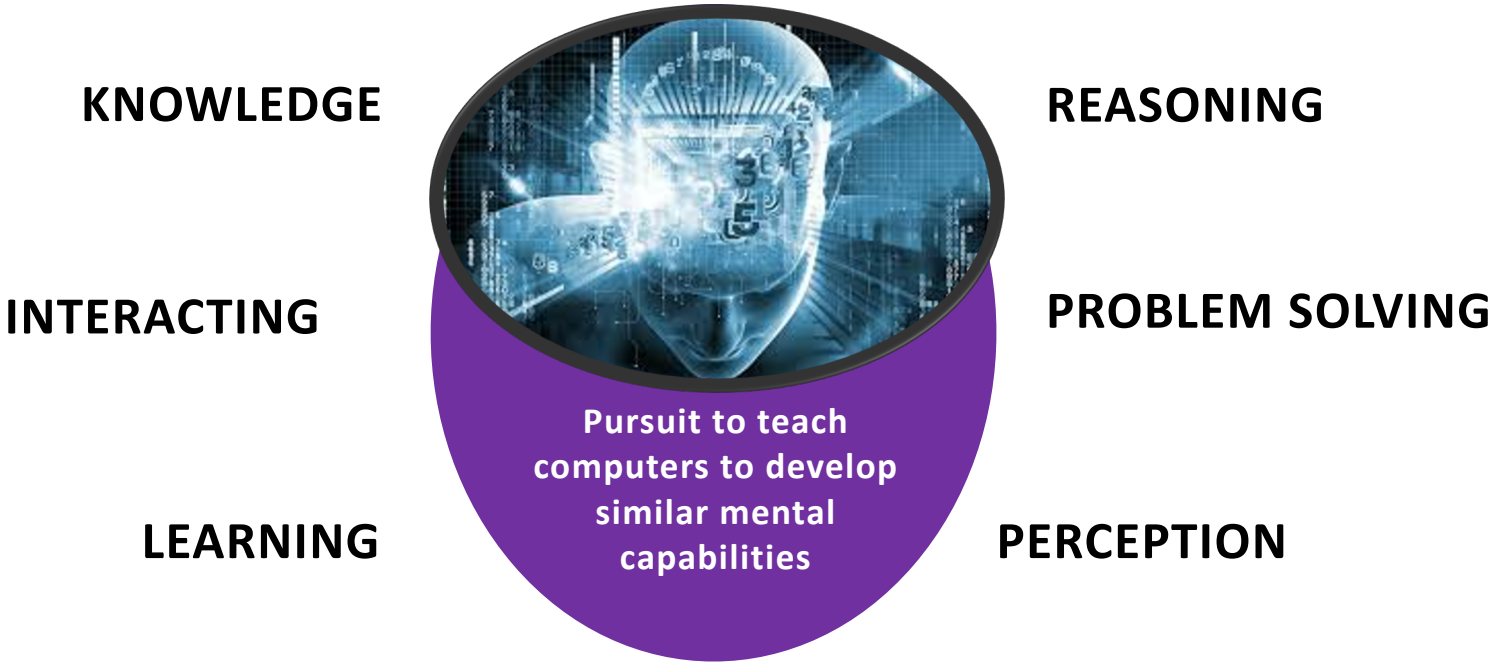
50,000
IT end users & storage pros worldwide

Agenda

- What is Intelligence, Artificial Intelligence and Machine Learning?
- The anatomy of an AI / Analytics Solution
- Building the AI Stack

What is intelligence?

- Intelligence is a person's mental capability to perceive, reason, act, learn quickly, and solve problems (among other things)

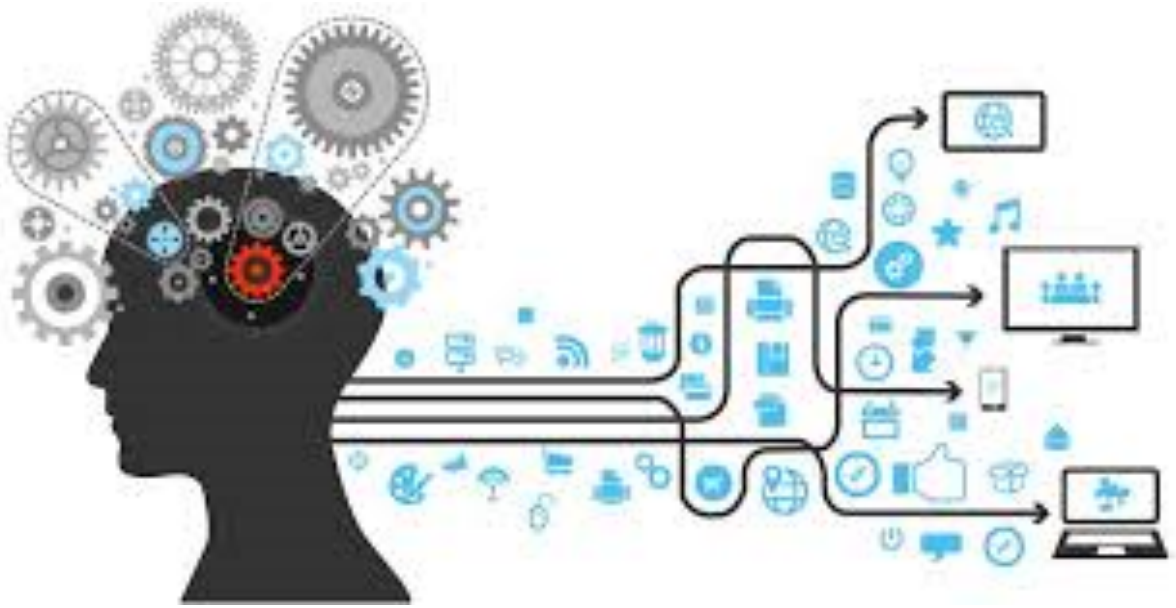


What is Artificial Intelligence?

- Definition
  - The ability of computer systems to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

– Example

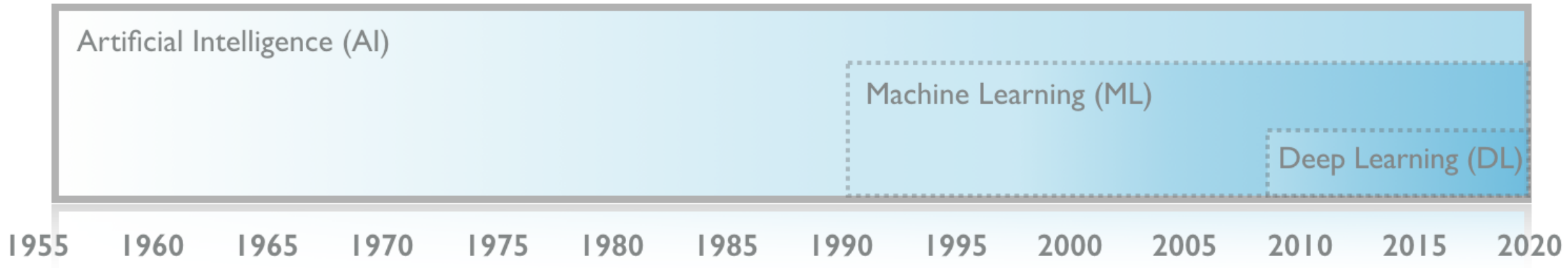
  – Turing Test: Inability to distinguish computer responses from human responses



History of AI

- AI has been around for more than 50 years
- The term AI was introduced in 1956 by John McCarthy, an American computer scientist
- The growth and adoption of Machine Learning and Deep Learning have made AI real

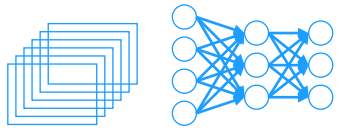
Evolution of Artificial Intelligence



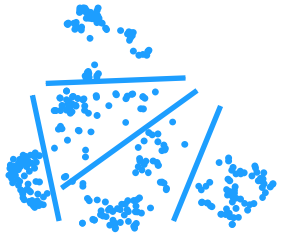
Many Approaches to Analytics & AI

No One size fits all…

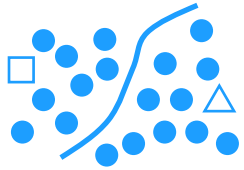
Supervised Learning



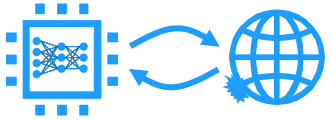
Unsupervised Learning

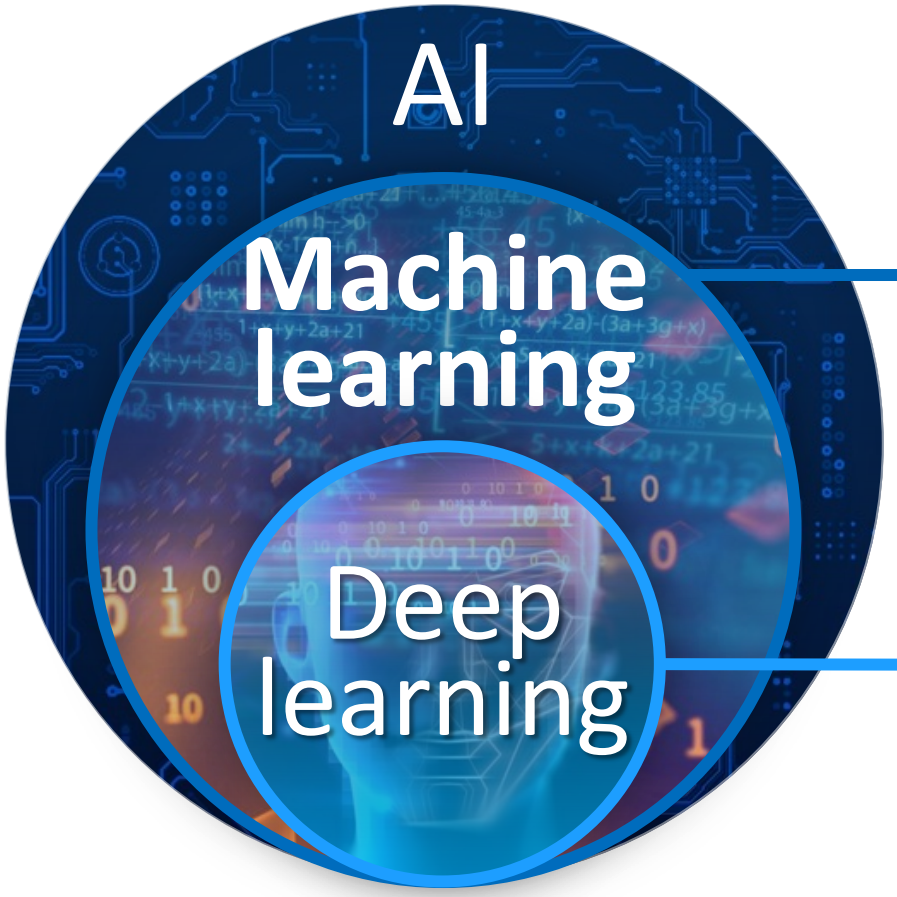


Semi-Supervised Learning



Reinforcement Learning





- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation

- Image Processing
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks

Why is Everyone Talking about AI Now?

Actionable data

Enormous amount of data of all types

Critical for machines to learn

Algorithms

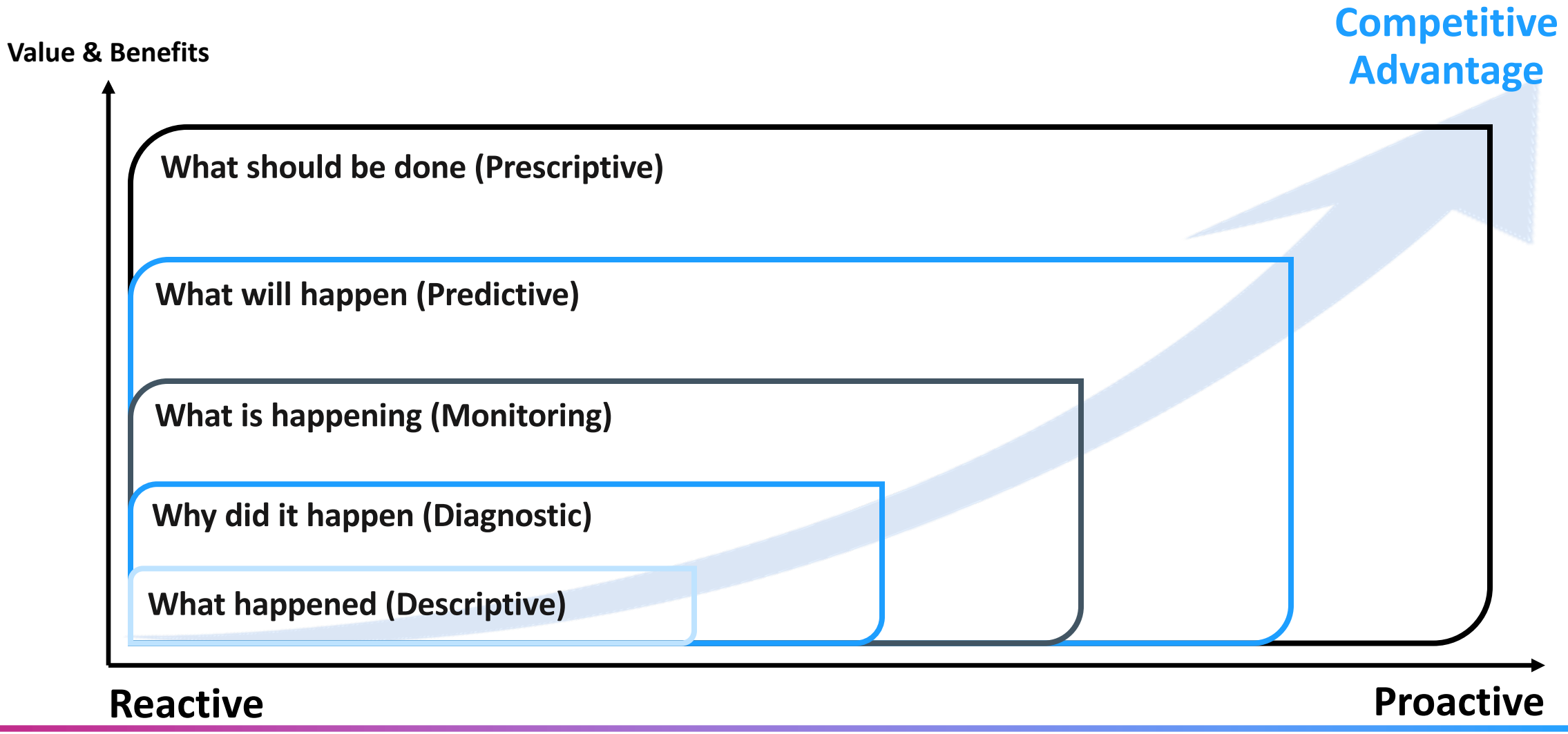
Availability of New Algorithms, Neural Network Research and Software

Compute power

GPU based computing with super compute power

Relatively low cost

From Descriptive to Prescriptive



Event Streams Simplify Data Pipelines

DATA PIPELINE OPTIMIZATION

Capture Your Data At or Near the Source

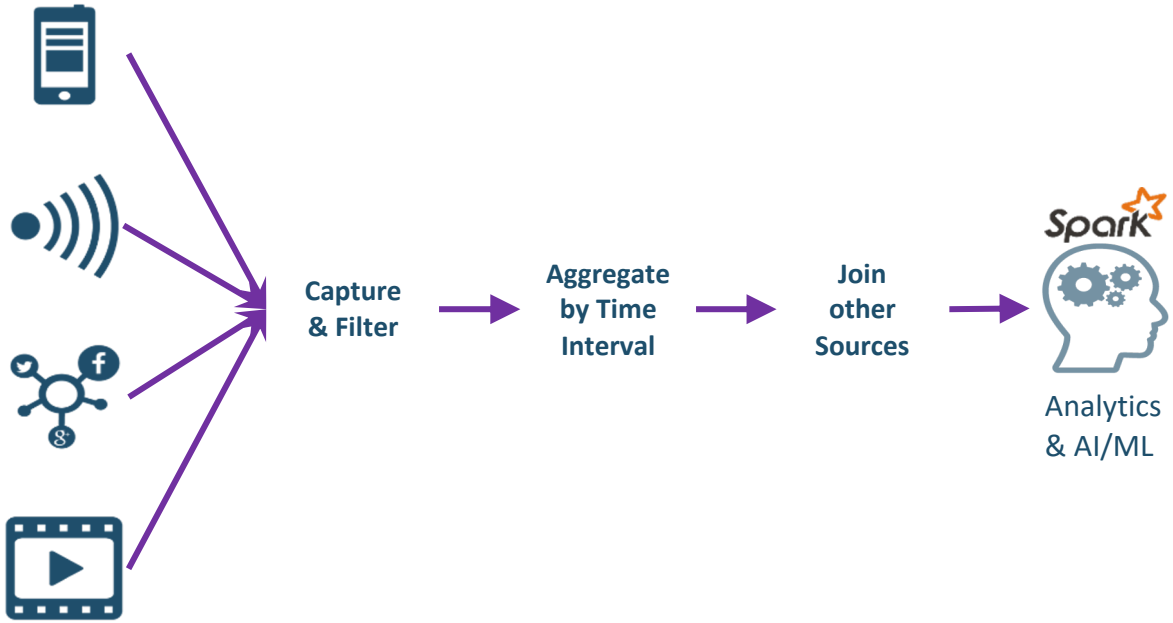
Filter for Specific Event Information

Take any **immediate actions** necessary

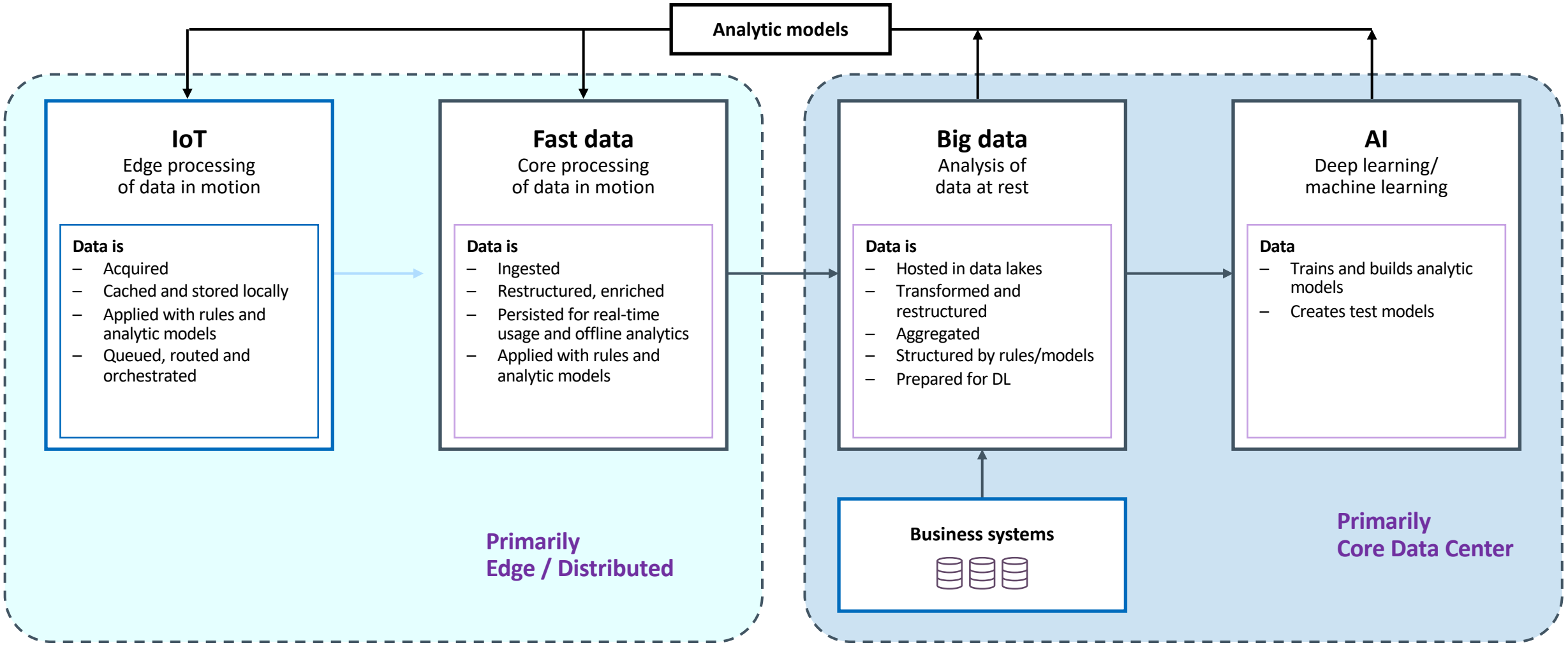
Aggregate By Time Interval

Join with Other Data Sources

Analyze with AI/ML, and Analytics Tools



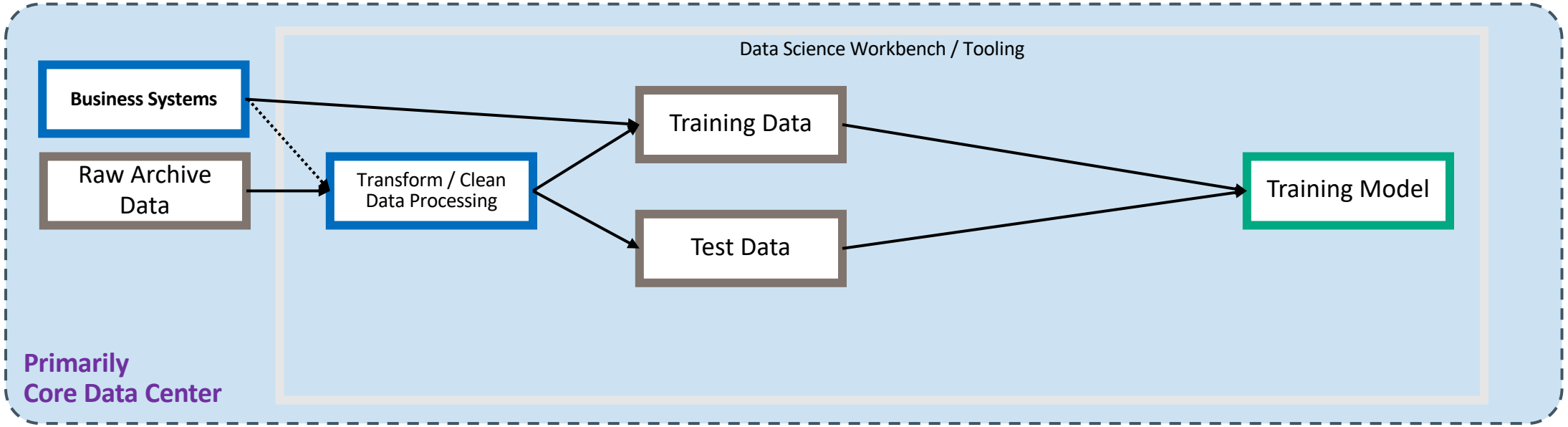
Unify End-to-End Data Pipeline for AI



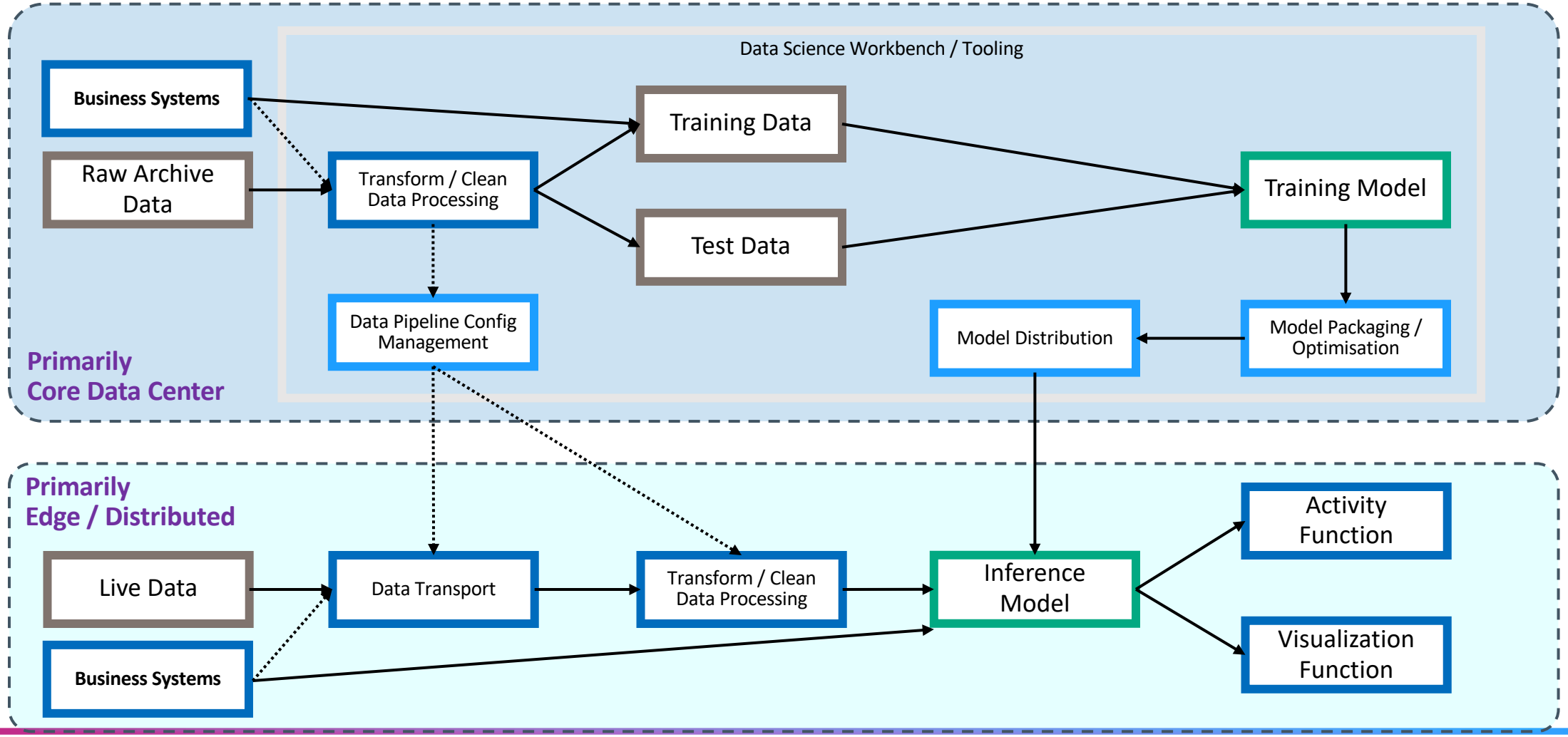
The Anatomy of AI Solutions

Training Model

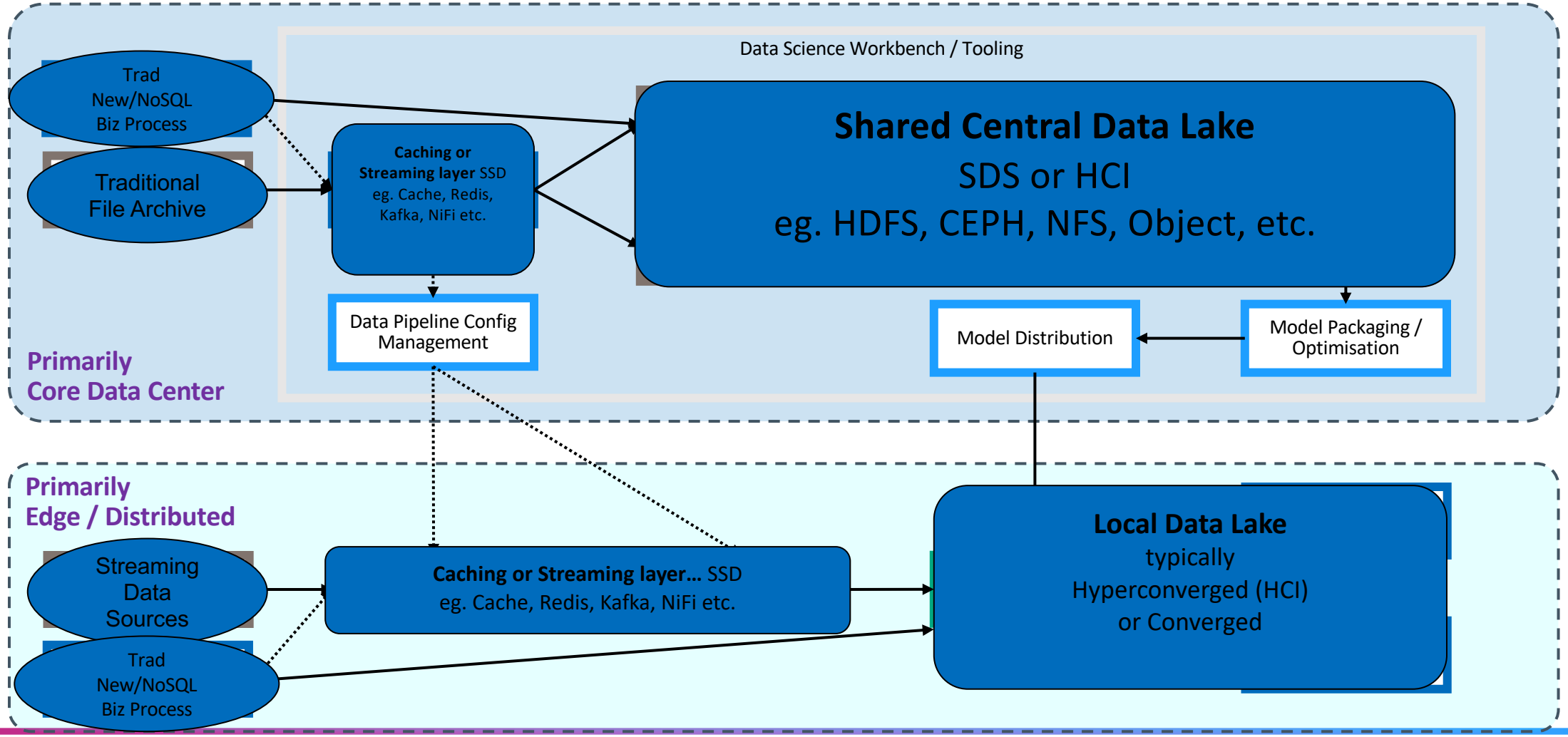
The Anatomy of AI Solutions



The Anatomy of AI Solutions

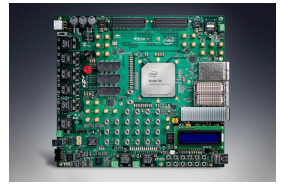


The Anatomy of AI Solutions



Building the AI Stack
Compute Hardware

- CPUs
  - Traditional source of raw compute
  - Used for both training & inference
  - Hybrids with other technologies
- GPUs
  - High speed floating point hardware
  - De facto for AI training
- ASICs & TPUs
  - Application Specific Integrated Circuit
  - Reduced precision FP (for training) or integer (for inference) operations
- FPGAs
  - Field Programmable Gated Arrays
  - Effective for inference
  - Reprogrammable



Software Frameworks

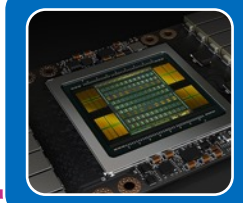
Frontend

- Abstracts the mathematical and algorithm implementation details of Neural Networks
- Provides a high level building blocks API to define neural network models over multiple backends
- A high level language library

Backend

- Hides hardware-specific programming APIs from user
- Optimizes and parallelizes the training and inference process to work efficiently on the hardware
- Makes it easier to preprocess and prepare data for training
- Supports multi-GPU, multi-node execution

AI Stack



Modern Compute

GPUs, TPUs, FPGAs

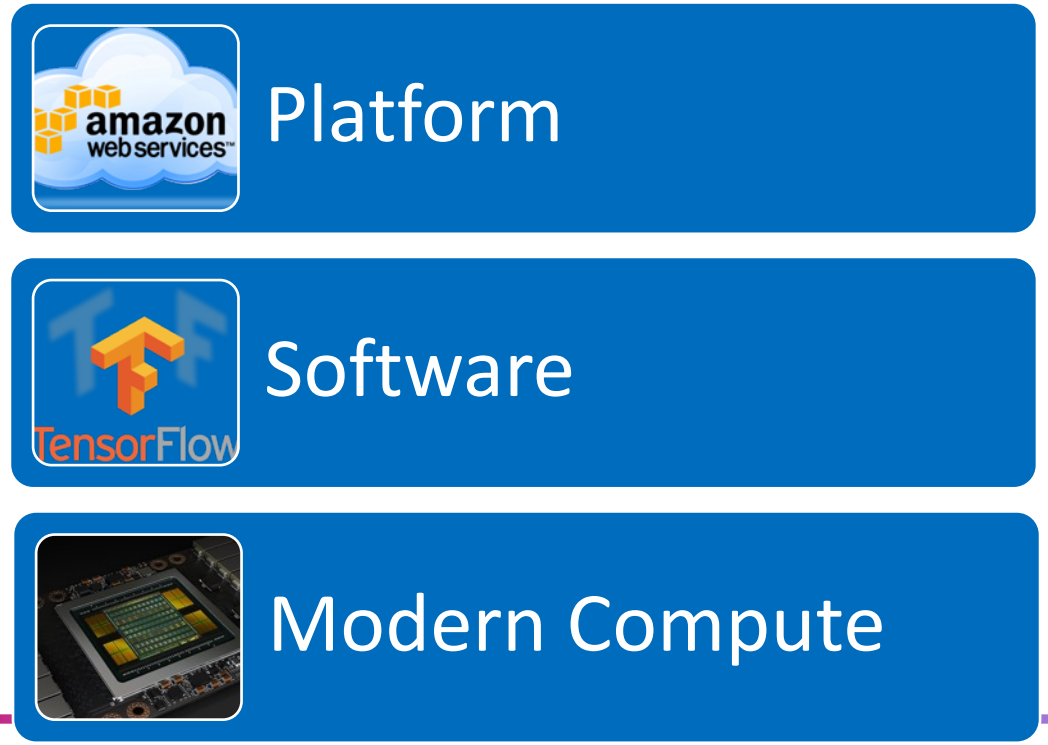
- Optimized hardware to provide tremendous speed-up for training, sometimes inference
- More easily available on cloud for rent

AI Stack

Tensorflow, PyTorch, Caffe2, MxNet, CNTK, Keras, Gluon

- Library that implements algorithms, provides execution engine and programming APIs
- Used to train and build sophisticated models, and to do predictions based on the trained model for new data

AI Stack



Laptop, Cloud compute instances, H2O Deep Water, Spark DL pipelines, HPE Container Platform & MLOps

- Hardware accelerated platforms, supporting common software frameworks, to run the training and/or inference of deep neural networks
- Typically optimized for a preferred software framework
- Can be hosted on-premises or cloud
- Also offered as fully-managed service (PaaS) by cloud vendors like Amazon SageMaker, Google Cloud ML, Azure ML

AI Stack

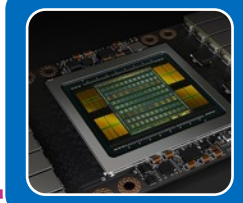
API-based service

amazon web services

Platform

TensorFlow

Software



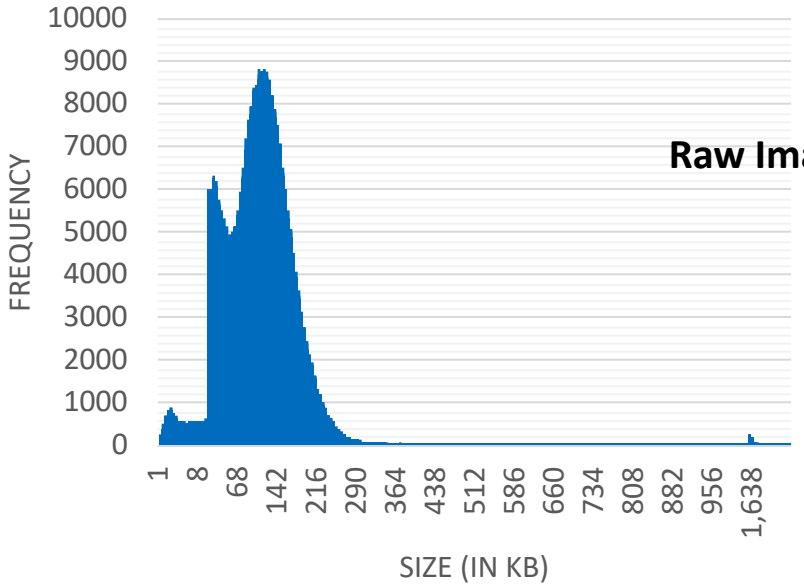
Modern Compute

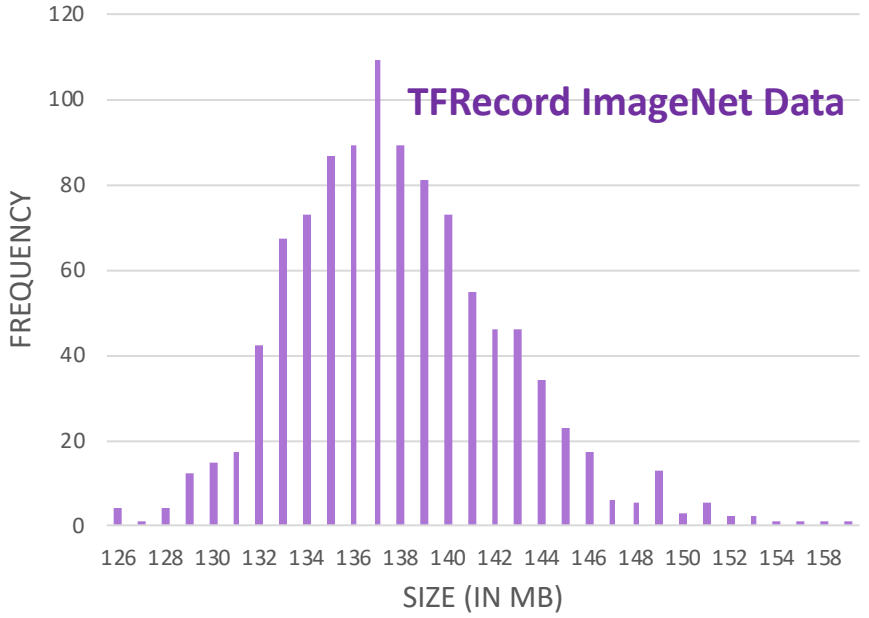
Amazon Rekognition, Lex & Polly; Google Cloud API; Microsoft Cognitive Services;

- Allows query based service access to generalizable state of art AI models for common tasks
  - Ex: send an image and get object tags as result, send mp3 and get converted text as result and so on
- No dataset, no training of model required by user
- Per-call cost model
- Integrated with cloud storage and/or bundled into end-to-end solutions and AI consultancy offerings like IBM Services Watson AI, ML & Cognitive consulting, Amazon's ML Solutions Lab, Google's Advanced Solutions Lab

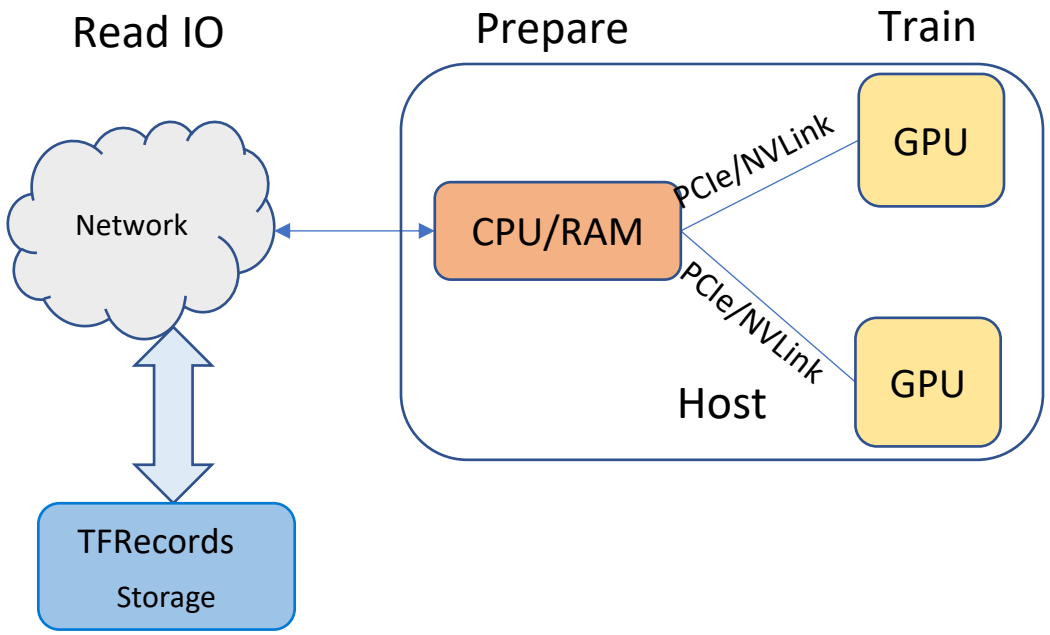
Dataset Transform – ImageNet Example

- Raw data vs TFRecords
- Raw data is converted into packed binary format for training called TFRecord (One time step)
  - 1.2 M image files are converted into 1024 TFRecords with each TFRecord 100s of MB in size





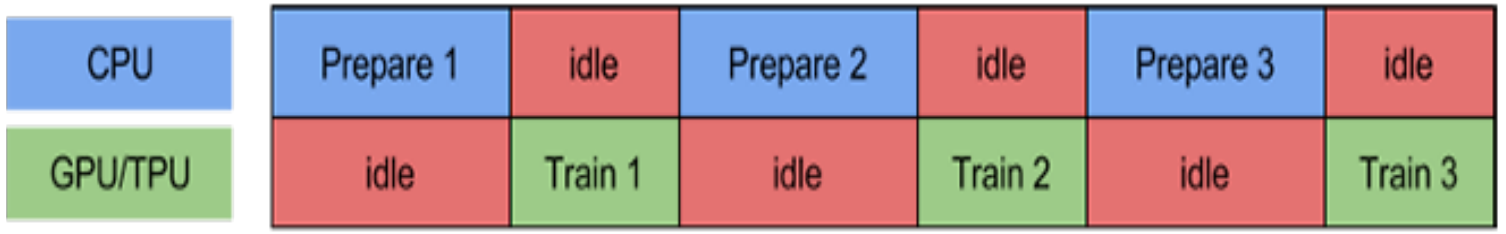
TensorFlow Data Pipeline



1. **IO**: Read data from persistent storage
2. **Prepare**: Use CPU cores to parse and preprocess data
   - Preprocessing includes Shuffling, data transformations, batching etc.
3. **Train**: Load the transformed data onto the accelerator devices (GPUs, TPUs) and execute the DL model

Compute Pipelining

- Without pipelining



- With Pipelining (using prefetch API)

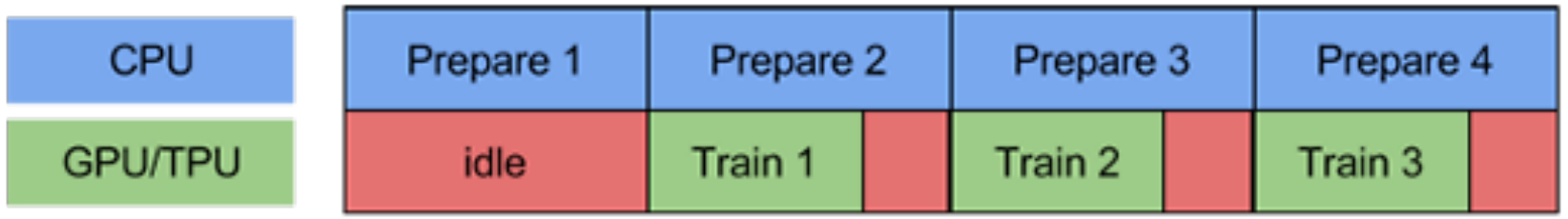
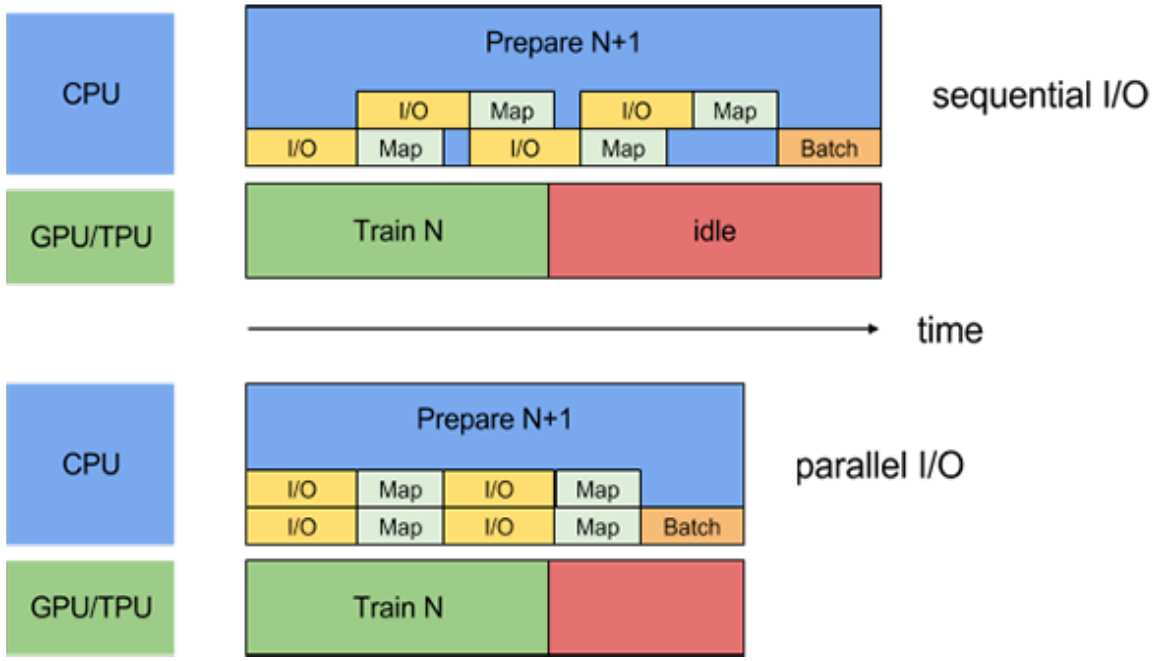


Image source: https://www.tensorflow.org/guide/

Parallelize IO and Prepare Phase

- Parallelize IO



- Parallelize prepare

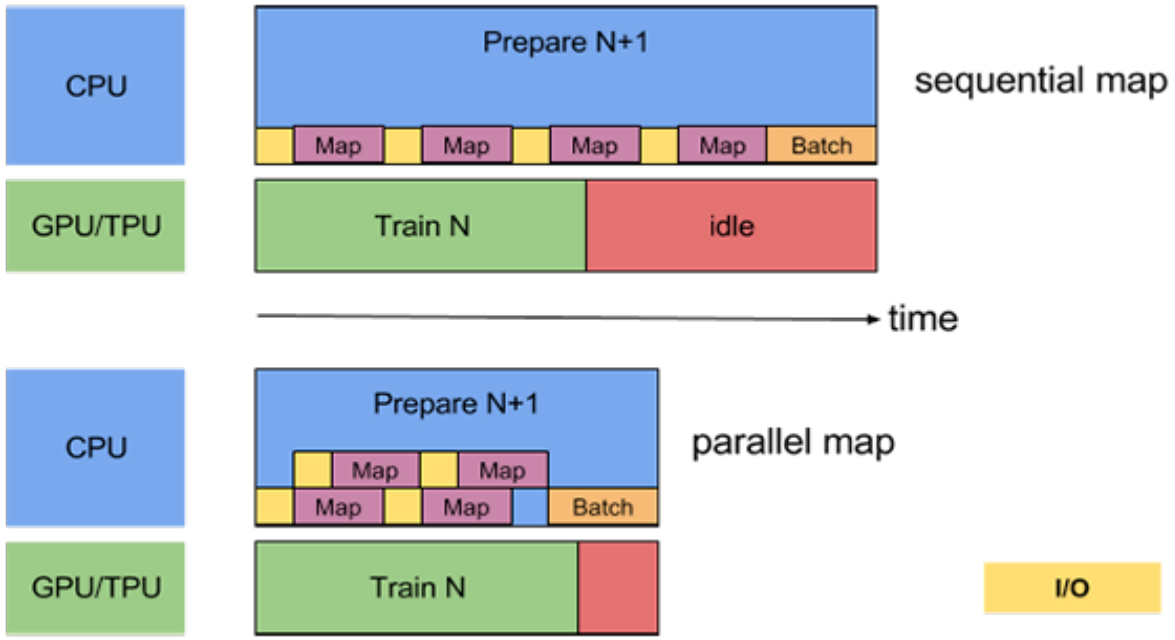


Image source: https://www.tensorflow.org/guide/

Research Directions in AI

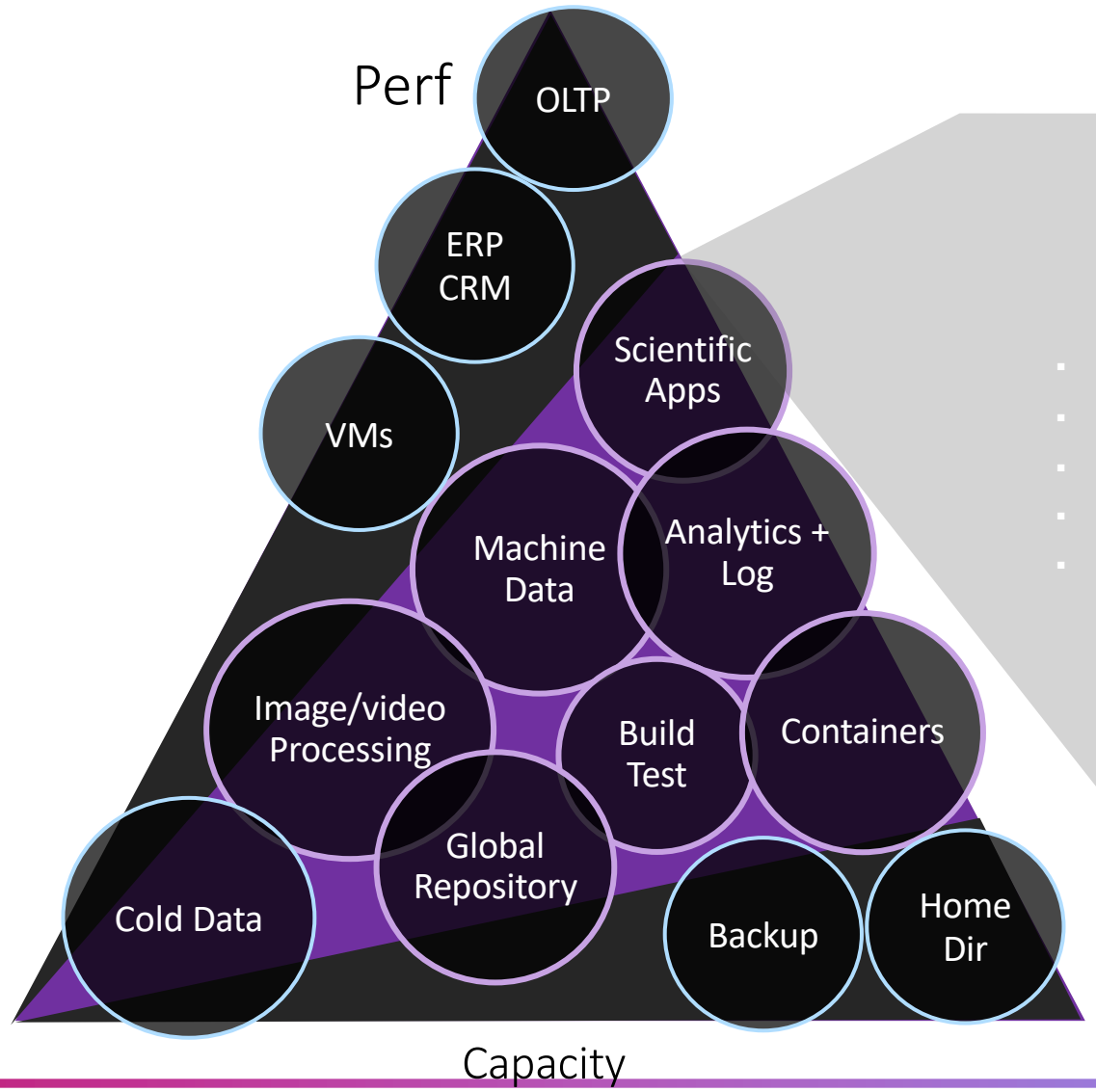
Academic:

- Using DL to replace heuristics-based decision within systems software, or even data structures
- Systems and platforms for DL
- Practical engineering optimizations to improve DL process/lifecycle/performance
- Workload and benchmarking
- Other areas like security, privacy, power etc.

Industry:

- Google Brain: hardware, AutoML
- FAIR: vision, video, AR
- Apple: speech & vision on-device

New Use Cases & Workloads Are Changing “Storage”



Workloads

- Legacy Operational apps
- Media & Entertainment
- Image Processing
- Analytics & Log Processing
- Container based apps

Capabilities

- Global Namespace
- Multi-temperature store
- Extreme Scale & Reliability
- Guaranteed Performance
- Agile Platform for heterogeneous data types
- Integrated Analytics

Delivered in a cloud-scale architecture with commodity economics across public and private clouds.

Summary

- AI is more than just training models!
- Data is no longer single purpose, but purpose+
- We need to simplify and unify data treatment
- Move to Shared Data Storage Architecture
- Unify Data Life Cycle
- Unify End-to-End Data Pipeline
- Pipelines are the new SAN

Additional Resources

- Presentation: Customer Support through Natural Language Processing and Machine Learning
  https://youtu.be/u1iRvWzMioM
- Presentation: Introducing the AI/ML and Genomics Workloads from the SPEC Storage Subcommittee
  https://youtu.be/47pmqFXYi-4
- White Paper: Is Your Storage Ready for AI?
  - https://www.intel.com/content/www/us/en/products/docs/storage/are-you-ready-for-ai-tech-brief.html
- Article: Want optimized AI? Rethink your storage infrastructure and data pipeline
  - https://venturebeat.com/2020/01/16/want-optimized-ai-rethink-your-storage-infrastructure-and-data-pipeline/

After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Cloud Storage Technologies Initiative website and available on-demand at https://www.snia.org/forum/csti/knowledge/webcasts
- A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud

Thank You!

