# ACCELERATED DATA OPERATIONS


**Author:**

**Shailesh Manjrekar**
*Head of Artificial Intelligence*
WekaIO

# INTRODUCTION

The unprecedented growth and sprawl of data results in a complex and varied ecosystem of management. The location of the data compounds the challenge of managing it. DataOps, or Data Operations, is a term probably first used by Lenny Liebmann, Contributing Editor of InformationWeek in a blog post in 2014. Gartner placed DataOps In its well-known Hype Cycle for Data Management, in 2018. Gartner defines DataOps as "a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization. The goal of DataOps is to deliver value faster by creating predictable delivery and change management of data, data models and related artifacts."

Data is the currency of today's economy. DataOps has risen as a discipline of connecting the data points from creators, to analyzers, to consumers and using data to create value, generate wealth, and deliver competitive differentiation. The role of DataOps will be a key differentiator for many enterprise organizations, as we desire to speed time-to-innovation and to-value, we need to accelerate DataOps.

# TABLE OF CONTENTS

# CHAPTER 1:
## It's All About the Data

Data has become the most important strategic asset to digital businesses for launching new business models, faster time to market, and competitive differentiation. Accelerated DataOps – Data Management in the AI era – determines how well businesses can derive actionable intelligence, operationalize pipelines, and provide governance and trust. It is going to determine success in the digital economy.

### New Workloads – Convergence of HPC, HPDA, and AI with Accelerated Computing

AI is penetrating the traditional High-Performance Computing (HPC) and High-Performance Data Analytics (HPDA) markets. For example, HPC customers are leveraging GPU-accelerated INDEX libraries for simulation on GPU compute, with WekaFS-enabled GPUDirect Storage. RAPIDS and BlazingSQL can run GPU-Accelerated Data Analytics and GPU-Accelerated SQL querying, and with the DALI library, run Deep Learning on GPUs. The Accelerated Compute layer is not just limited to GPUs but involves FPGAs, Graph processors, and specialized accelerators.

The use cases are moving from computer vision to Conversational AI, NLP / NLU, and multi-modal use cases. Recommendation engines from Alibaba and Baidu are now using deep learning, while you see low latency inference used for personalization (LinkedIn), speech (iFLyTek), translation (Google), and video (YouTube) use cases.

Learning (Training) is quickly moving from supervised (labeled data) to using Convolution Neural Networks (CNNs) for annotation and labeling, to transfer learning (where one Deep Neural Network (DNN), trained on one set of datasets, can also be trained for other datasets), to federated learning (centralized), to active learning (label selective scenarios). Similarly, DNNs are becoming more and more complex, as with the several billion hyper-parameters of BERT and Megatron.

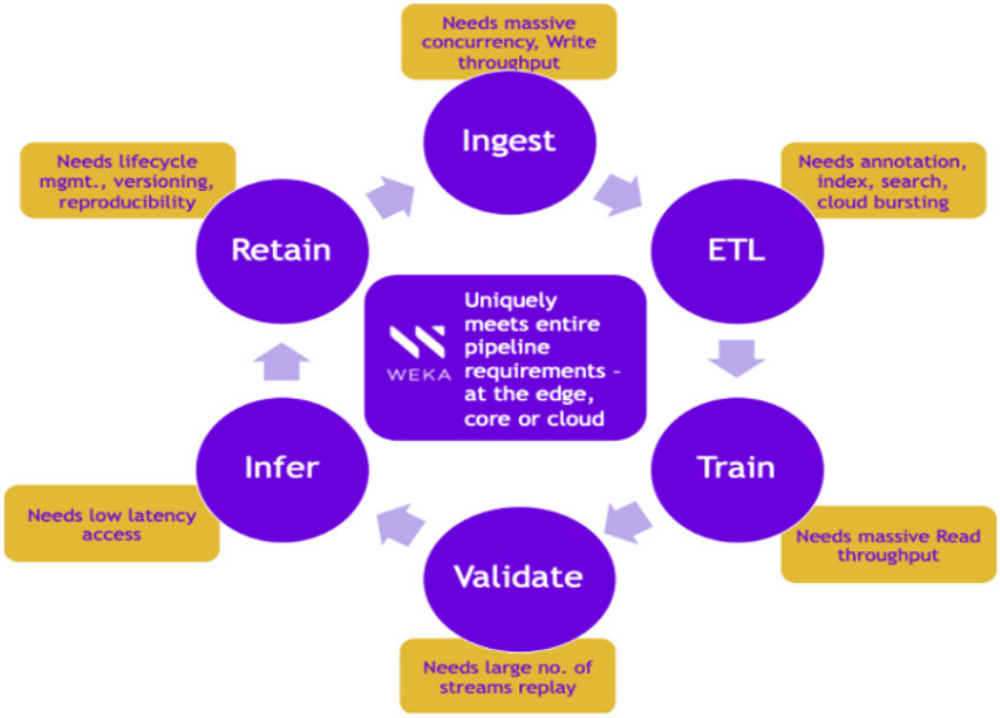


*Figure 1: The stages in the AI data pipeline have distinct data requirements.*

All of these transitions result in different stages within AI data pipelines that have distinct data (storage and I/O) requirements for massive ingest bandwidth, mixed read/write handling, and ultra-low latency, often resulting in a storage silo for each stage [Figure 1]. This means that business and IT leaders must reconsider how they architect their storage stacks and make purchasing decisions for enabling Accelerated DataOps.

# CHAPTER 2:

## New Workloads Drive the Need for Modern Underlying Architectures

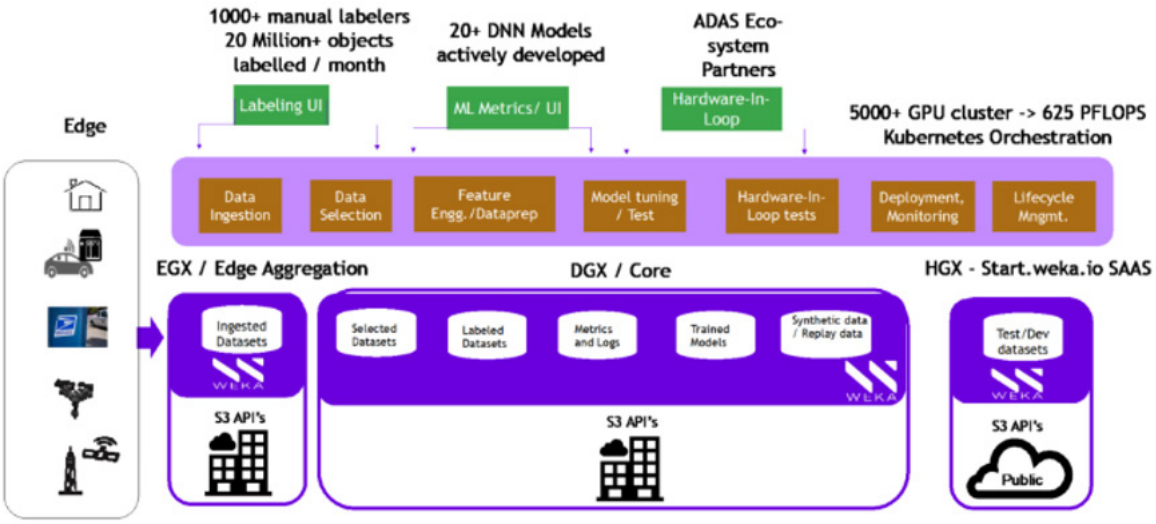


*Figure 2: New Edge AI-to-Core-to-Cloud architectures*

With the advent of IoT, 5G, and low-powered devices, Edge AI is expected to be bigger than the cloud. Depending on the use case, an Edge endpoint could be an autonomous vehicle, a drone, a Point of Sale device, an IP camera, or a precision medicine device. There is often an Edge aggregation point that is increasingly seen in parking lots, post offices, retail stores, and Edge data centers. This Edge endpoint aggregates the Edge datasets and runs Edge-to-Core-to-Cloud data pipelines. As a result, the infrastructure increasingly needs to cater to Edge-to-Core-to-Cloud data pipelines.

Also, in order to cater to complex DNNs and the convergence of HPC, HPDA, and AI, architectures such as GPUDirect storage are becoming paramount to feeding the GPU memory directly, providing the highest bandwidth and lowest latencies. Sharing datasets for distributed training across 64 DGX-2s (5,000 cores x 16 Teslas x 64 DGX-2s) has become the norm. Imagine the parallelism that is involved at the compute layer. Transports such as NVMeOF (over InfiniBand or RoCE Fabrics) are making data locality a non-issue, especially with support for 100 Gb/sec and 200 Gb/sec networking.

Equally important is the ability to cater to performance at scale, with use cases such as ADAS. A single survey car being used to achieve a goal of SAE 2+ autonomy, with 8*2 MP cameras and 10 DNNs, easily generates 2 PB of datasets per year.

Storage architectures using traditional direct-attached NVMe storage (DAS), NAS, and object storage limit performance and data mobility. In addition, NVMe block solutions lack the shareability and parallelism to deliver timely insights at scale for these new workloads and architectures.

## Digital Transformation with Accelerated DataOps – Business and IT Convergence

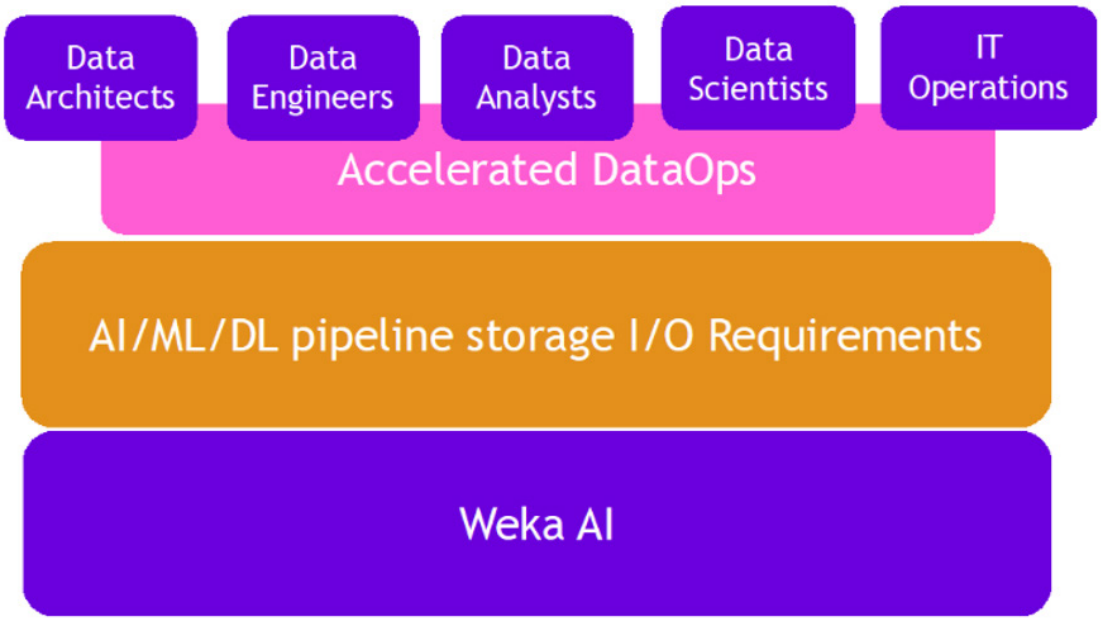


*Figure 3: DataOps line of business personas*

When it comes to new workloads and architecture, these projects typically originate with Line of Business personas – Chief Data Officers, Chief Analytics Officers, CIOs, Data Scientists, Data Engineers – and then are augmented by IT personas. It is important to have an IT infrastructure, and particularly a storage solution, that can cater to both Line of Business and IT requirements. If not, the result will be data and IT silos. Accelerated DataOps break these silos by enabling shared goals, vision, and infrastructure. Weka AI enables Accelerated DataOps as follows:

- **Accelerated DataOps for Analytics – derive actionable intelligence**
  - Run Business Intelligence with GPU-accelerated libraries such as RAPIDS and query engines such as BlazingSQL while running Artificial Intelligence with DALI on the same storage substrate
  - Run Descriptive, Predictive (what-if) and Prescriptive (what-if-then) and Cognitive Analytics with same storage substrate
- **Accelerated DataOps for Operational Agility – improve productivity, reduce TCO**
  - Data is new source code – Weka AI provides data versioning, test/dev for pipelines, data protection with backup and recovery and DR
  - Data Anywhere – Weka AI manages Edge-to-Core-to-Cloud pipelines
  - CloudStore – Weka AI manages performance (NVMe Flash) and capacity tiers (S3/HDD) as a single namespace. Weka AI supports a broad ecosystem of S3 on-prem and public object stores
- **Accelerated DataOps for Governance**
  - Weka AI provides line-rate in-flight and at-rest encryption supported by popular Key Management Systems such as HashiCorp Vault and supports virtual filesystems.

# CHAPTER 3:

## Weka AI for Accelerated DataOps

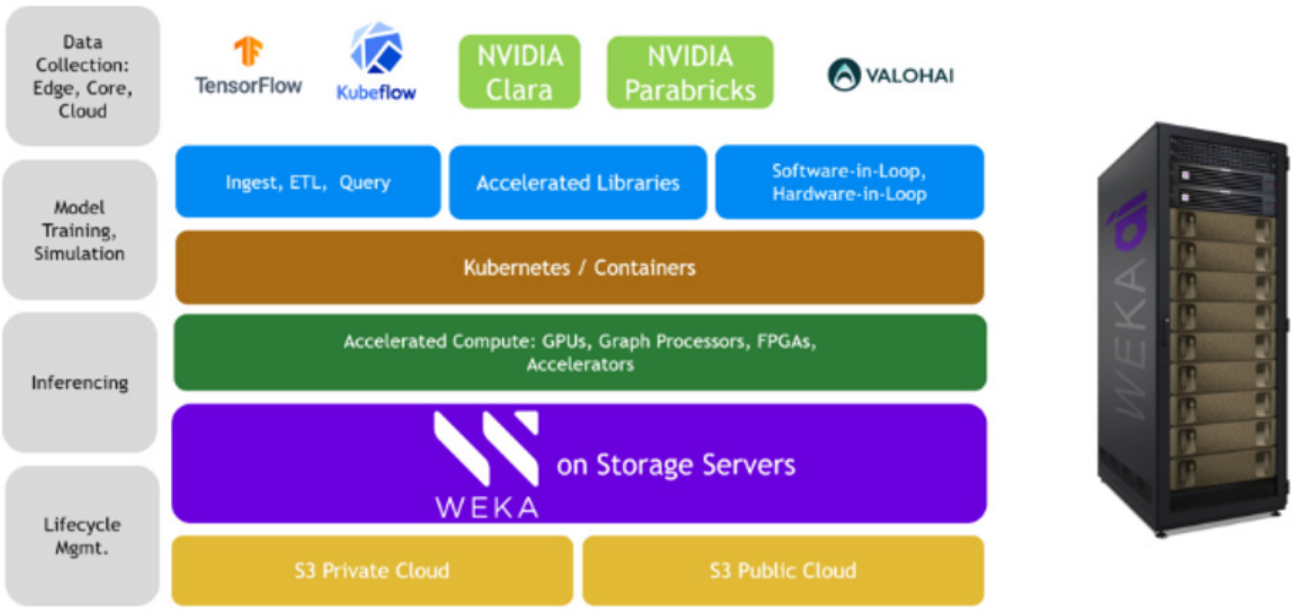


*Figure 4: Weka AI is architected to enable Accelerated DataOps*

Weka AI is architected to enable Accelerated DataOps by solving these storage challenges and delivering production-ready solutions with Reference Architectures and Software Development Kits (SDKs). Weka AI empowers Accelerated DataOps by breaking storage silos, enabling convergence across HPC, HPDA and AI workloads, Business Intelligence (BI), and Artificial Intelligence (AI) on the same storage substrate. Weka AI delivers operational agility with versioning, explainability, reproducibility, governance, and compliance with in-line encryption and data protection. Working with technology alliance partners, Weka AI provides a production-ready solution where the entire AI data pipeline workflow — from data ingestion to batch feature extraction, to training, to hyperparameter optimizations, and finally to inference and versioning — can be run on the same storage platform, whether running on-premises or in the public cloud.

### Explainable AI (XAI) – Integration with Valohai's Deep Learning Pipeline Management

Explainable AI is paramount when it comes to use cases such as autonomous driving, healthcare, and genomics as they have social impact. Deep Neural Networks (DNNs) for the most part are black boxes comprised of several hidden layers. The only way an experiment can be explained is by examining the dataset that it was trained on, audit trails, and lineage.

Weka demonstrates this with our integration with Valohai – a Deep Learning pipeline management system. This demo outlines how Valohai and WekaFS are integrated in an AWS Virtual Private Cloud (VPC) to run a pipeline for image classification, using the popular CIFAR-10 database and TensorFlow model. Data scientists can use the popular Jupyter Notebook or the Valohai GUI for pipelines to do:

1. data transformation and model training
2. hyperparameter optimization and finally
3. inference.

Valohai DLMS seamlessly integrates Weka's powerful snap2object capabilities, where the data science experiment is version-controlled with the code, data, audit logs, and lineage and can be easily reproduced and explained whenever needed. Additionally, in-line encryption capabilities of Weka, with leading Key Management Systems such as HashiCorp Vault, provide data security, compliance, and governance.

## Solution Reference Architectures Powered by Weka AI

Weka AI is based on proven deployments with customers and Weka Innovation Network (WIN) Partners and addresses several vertical use cases:

- ADAS – semantic segmentation for annotating Automated Driver Assistance System datasets
- Deep Learning pipeline management solution with Valohai
- Life Sciences – next-generation sequencing solution with Parabricks and HPE
- Healthcare – integrated medical imaging solution with NVIDIA Clara
- FSI – STAC M3 testing with Kx Systems and HPE
- Retail – RAPIDS and BlazingSQL-based solution, fraud analytics
- Oil and Gas – HPE AI DataNode with Weka and Scality
- WeKa AI Reference Architecture for training and inferencing
- Public sector HPC solution with Penguin

**Weka AI benefits new personas as follows:**

**Chief Data Officers (CDOs), Chief Analytics Officers (CAOs), and Line of Business Data Scientists**

- Reducing epoch times from days to hours, while delivering the lowest inferencing times and maintaining the highest images/sec benchmarks. This is enabled by industry-best GPUDirect storage performance of 80 GB/sec to a single DGX-2 client
- Explainability and reproducibility for experiments using instant, space-efficient snapshots
- Hybrid workflows – Dev and Test experiments in the public cloud and seamless movement to on-premise or production
- Data Compliance and Governance with in-flight and at-rest encryption

**Data Engineers and IT Leaders**

- Best TCO by leveraging NVMe flash for performance and HDD object for capacity, with built-in data protection
- Eliminating silos and multiple copies, but providing a single storage platform for the entire data pipeline
- Best agility with data management across the edge, core, and cloud
- Best scalability with up to EBs of storage with trillions of files across directories and billions in a single directory
- Ease of management, a single point for support, and easy-to-consume as small, medium, and large bundles

# SUMMARY

Weka offers WekaFS, the modern file system that uniquely empowers organizations to solve the newest, biggest problems holding back innovation. Optimized for NVMe and the hybrid cloud, Weka handles the most demanding storage challenges in the most data-intensive technical computing environments, delivering truly epic performance at any scale. Its modern architecture unlocks the full capabilities of today's data center, allowing businesses to maximize the value of their high-powered IT investments. Weka helps industry leaders reach breakthrough innovations and solve previously unsolvable problems.

WekaIO™ (Weka) is uniquely positioned to anticipate the needs from AI market transitions and provide transformative solutions, like Weka AI. Weka, through delivering these solutions, makes it easy for our customers to monetize their data, achieve faster time-to-market, and gain competitive differentiation with the best TCO. Weka AI is a transformative solution framework that caters to these market transitions and delivers compelling benefits to Line of Business personas.

## Additional resources:

- Valohai and Weka demonstration [video]
- WekaFS for AI and Analytics
- Weka AI Reference Architecture
- Gartner Glossary: DataOps
- Lenny Liebmann, Contributing Editor, InformationWeek, DataOps blog post, 2014

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io **www.weka.io**

©2017-2020 All rights reserved. WekaIO, WekaFS, Weka AI, WIN, Weka Innovation Network, the Weka brand mark, WIN logo, Weka AI logo, Weka logo, and Radically Simple Storage are trademarks of WekaIO, Inc. and its affiliates in the United States and/or other countries. Other trademarks are the property of their respective companies. References in this publication to WekaIO's products, programs, or services do not imply that WekaIO intends to make these available in all countries in which it operates. Product specifications provided are sample specifications and do not constitute a warranty. Information is true as of the date of publication and is subject to change. Actual specifications for unique part numbers may vary.

W01EB202008