Accelerated DataOps-Data Management in an AI Era

Accelerated DataOps—Data Management in an AI Era
Tom Coughlin, Coughlin Associates, Inc.

The growth of IoT sensor and camera networks, higher resolution and higher frame rate video as well as the wide spread applications of various artificial intelligence (AI) tools are driving an explosion of data at the edge, in the network core and in data centers. Much of that data has immediate value and must be processed and interpreted quickly. Processing this explosion of data has increased the generation of value in our world and is driving the digital economy.

Data has been rightfully called the new "black gold." Like oil in the earlier industrial age, data is driving and enabling modern businesses. The large pools of data that organizations are collecting are often referred to as data lakes. The need to discover and use data to create new value and generate wealth with faster time-to-market and competitive differentiation, leads to new ways to find, manage and use data in these data lakes, referred to as DataOps.

DataOps
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 ln 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."

Accelerated DataOps
Accelerated DataOps focuses on more efficient and timely management and application of data to solve real-world problems using AI tools. [Accelerated DataOps]() creates actionable intelligence that allows users to further their success in the digital economy. Accelerated DataOps works by solving challenges associated with the use of digital storage, operational workflows and system architecture.

In practice, accelerated DataOps needs to:
1. Provide actionable management of business intelligence and artificial intelligence, while catering to a multitude of diverse I/O requirements, at the edge, the core and the cloud.
2. Provide operational agility for continuous improvement and continuous development (CI/CD), whether on- premises or in the cloud.
3. Provide end-to-end governance and security for data in-flight and data at rest.

AI applications are causing the convergence of high-performance computing (HPC) and high-performance data analytics. Both operations are being run on the same accelerated, massively parallel and increasingly distributed computer platform. These compute platforms include various special purpose (specific domain) computing devices such as GPUs, FPGAs, and various other specialized processors. These AI applications are enabling more rapid business decisions, based upon faster analysis of data. This is Accelerated DataOps in action.

Accelerated DataOps enables fast decision making using business intelligence (BI) and AI applications and includes descriptive, predictive, prescriptive and cognitive analytics that function across a unified storage pool that spans from the edge to the core and to the cloud. Accelerated DataOps for operational agility can improve business productivity and reduce the system total cost of ownership (TCO) by easy access and analysis of data in any location, managing performance and capacity storage tiers including cloud and multi-cloud. Accelerated DataOps should also provide good governance capabilities such as in-line encryption and virtual filesystems.

The figure below shows how various users will use accelerated DataOps for accessing a unified storage pool with the help of WekaIO (Weka). A single storage layer (single namespace) helps to break down silos between Line of Business and Infrastructure administrations.
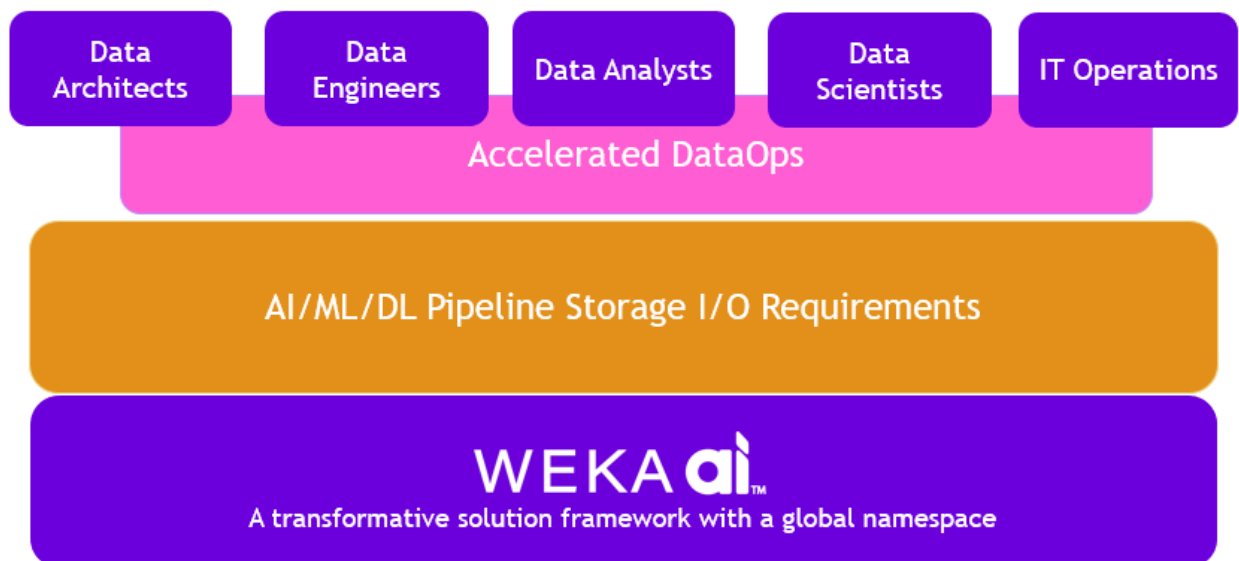
Weka says that their WekaFS, the world's fastest exascale filesystem, enables Accelerated DataOps that helps companies eliminate storage silos and infrastructure sprawl, while deriving actionable insights. Weka provides a unified storage substrate (or a unified data lake with a single namespace) using lots of parallel processes, with distributed meta-data; that can run at the network edge, core and in large data centers. Weka can utilize NVMe over fabric (NVMe-oF) for GPU accelerated data analysis as well as hard disk-based object storage and with both on premises, cloud and multi-cloud content. Working with NVIDIA's GPUDirect and an NVMe-oF storage system, over 82 GB/s performance was demonstrated on a single DGX-2 GPU server. For NFS mount points a 75X performance improvement can be achieved. Weka also excels at mixed workload performance, due to its architecture which unlike traditional storage, is optimized for both small and large random, sequential and metadata I/O's.

Weka works with the object storage ecosystem as its backend and has many partners. Rather than effectively tiering data it extends file name spaces to extend over object stores as well as

NAS storage, creating a unified file-based access to private and public object storage with accelerated analytics and access to data enabled by AI and workload optimization to match application requirements. Weka also provides persistence volumes orchestrated with Kubernetes for stateful AI/ML/DK data pipelines. The image below shows a Weka deployment to create a global namespace across multiple stored information sources.
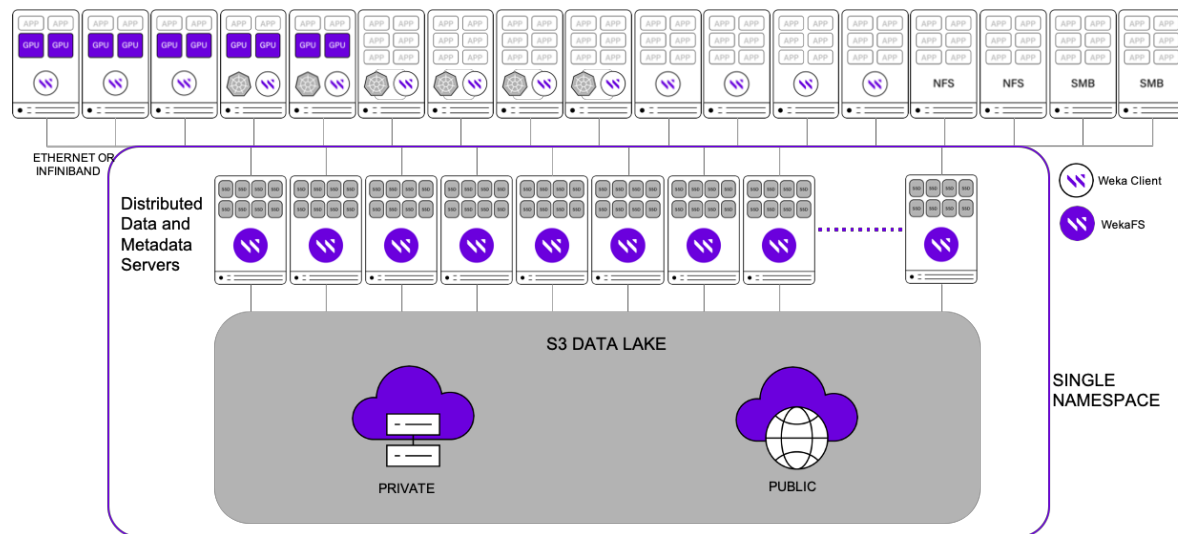
Weka provides Accelerated DataOps for small, medium and large data collections. The company also partners with enterprise providers such as Hitachi Vantara, who announced that Weka's scalable high-performance file system can be used to enhance Hitachi Vantara's NVME-native parallel file system network attached storage (NAS) HCP datastore for use in artificial intelligence, machine learning and analytics applications across a broad array of industries.

Hitachi Vantara said that the new OEM relationship with WekaIO will enhance Hitachi Vantara's portfolio with a high-performance, NVMe-native, parallel file system that the company will deliver tightly coupled to an HCP datastore. This performant network-attached storage (NAS) solution will be well suited for use with artificial intelligence, machine learning and analytics applications across a broad array of industries.

Accelerated DataOps Case Studies
Numerous organizations are benefiting from accelerated DataOps using Weka's unified file-based access to public and private object storage. These organizations include **TuSimple**, based in San Diego, which is developing deep learning for autonomous transportation fleets. While in the past **TuSimple** suffered from no data protection, multiple storage solutions, and long training times, with the Weka solution three clusters with 500 TB of SSDs and 1.5 PB of object storage increased the speed and quality of training with a fully protected data set and at less cost than an all-flash solution. The image below shows a **TuSimple** autonomous truck.

Cerence is a company that is building speech language models to improve the in-car experience. They are using AI to deliver a deep understanding of human behavior, culture and language in order to make automotive transport safer and provide more natural and comfortable human interactions with a vehicle. They needed a cost-optimized way to bridge from local file access to object storage and the public cloud to achieve scalable storage. The figure below is a schematic of the Cerence solution using a Weka file system.
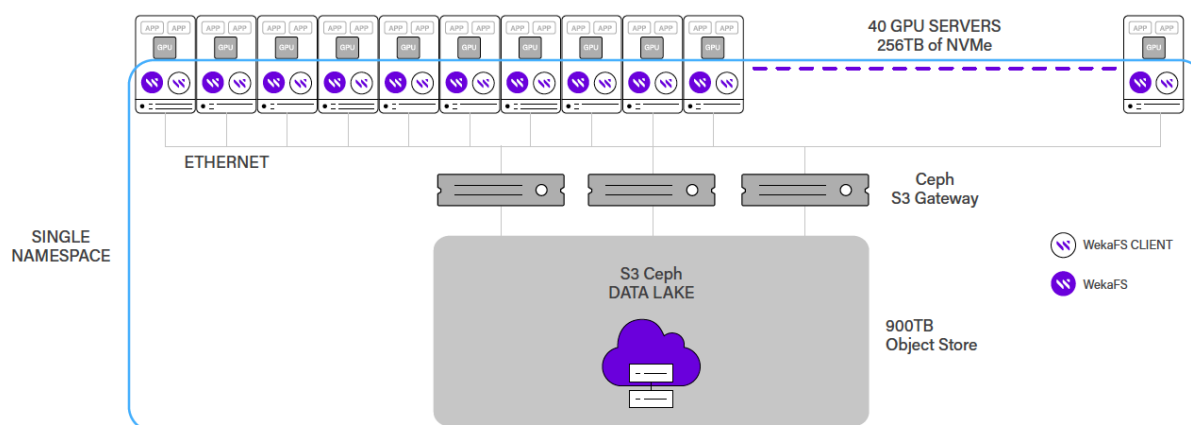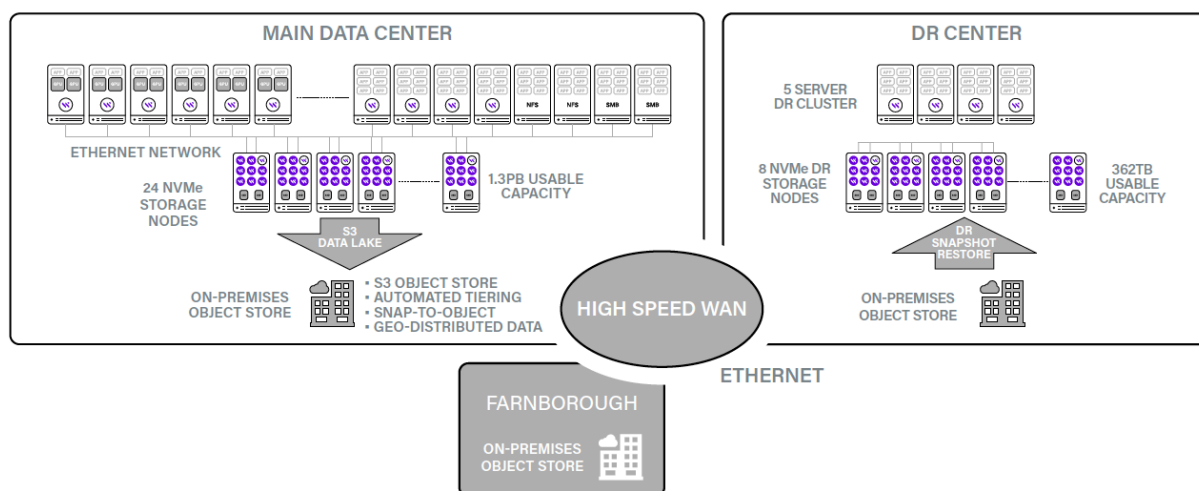

The Weka two-tier architecture using NVMe SSDs and HDD-based object storage presented as a single hybrid storage solution provided what Cerence required. The Cerence team managed 900 TB of data on the Weka file system to support NLP and NLU workloads on a cluster consisting of 40 HPE Proliant DL360 servers, each with dual 25GbE networking adapters.

The servers are interconnected with 4 switches for high availability (HA), performance, and redundancy. Each server has one network connection and two NVMe drives dedicated to WekaFS and one GPU card. Cerence IT is managing 3.2PB of data on object storage with SUSE Enterprise

Storage^{TM}, with 900TB assigned to Weka, running on a cluster consisting of 9 HPE Apollo 4200 servers, each with twenty-four 14TB drives. The team also utilizes an HPE Apollo 6500 server with 8 GPUs for multi-GPU processing.

Genomics England (GEL) is headquartered in London, England and is part of the UK Department of health and Social Care. The 5 million genomes project, which it announced in 2018, aims to sequence 5 million genomes from UK National Health Service patients with rare diseases. This data is then used for medical research. After completing the sequencing of 100,000 genomes, resulting in 21 PB of genome data, GEL projected that it would need 140 PB by 2023 to continue its work. To be useful researchers must have access to the entire data set and be able to access the data randomly.

GEL found that the scale-out NAS solution it used for the 100,000 genome sequencing had hit its limit on capacity scaling and had no viable disaster recovery strategy. It determined that for its 2023 growth plans it needed 140 PB in a single storage system, the storage system needed to provide very high performance to support research and access needs, it must also provide a disaster recovery solution, be easy to manage and fit within an allocated budget. The figure below shows how GEL used Weka to provide their desired storage solution.

WekaFS delivered a two-tier architecture that takes commodity flash and disk-based technologies and presents it as a single hybrid storage solution. The primary tier consists of 1.3 Petabytes of high performing NVMe-based flash storage which supports the working data sets. The secondary tier consists of 40 Petabytes of object storage to provide a long-term data lake and repository. Weka presents the entire 41 Petabytes as a single namespace.

Each of the tiers can scale independently: should GEL require more performance on the primary tier, it can do so independently of the data lake. The system takes advantage of the geo-distributed capability of the object store, and data is protected across three locations, 50 miles apart. If there is a major disaster in the primary location, Weka's unique Snap-to-object feature

allows the system to be re-started in a second location, ensuring continued access to the data. The geo-distribution and object store resiliency provide significantly higher data protection, as full back-up is impractical at this scale.

As a result of moving to this new storage system GEL has a 10X+ increase in performance compared to their prior NFS-based NAS. The system delivers over 135 GB/s from the NVMe tier, with performance scaling as the cluster grows. Also, the costs per genome dropped by 75% in storage costs and should achieve a 96% cost reduction by 2023. GEL can survive a major disaster at the primary site and still maintain access to the data. In addition, critical data sets are fully encrypted from the high-performance compute cluster all the way to the permanent data store and the system is protected from rogue security threats through a robust authentication mechanism.

Summary

DataOps has become an important element in modern data management and for today's big data workloads, accelerated DataOps are needed that can cost effectively process large amounts of data as fast as possible. WekaIO provides a unified file system that encompasses NAS and object storage to deliver high performance storage with modern data protection and cost-effective tiering. The Weka solution has benefited many end users; from autonomous vehicles, to speech recognition, financial services industry (FSI) and genomics research.

Additional resources:
- [TuSimple Selects Weka for Autonomous Fleet Vehicle Machine Learning TrainingCerence case study](#)
- [GEL Reduces Research Cost-Cost per Genome-With WekaFS](#)

About the Author

Tom Coughlin, President, Coughlin Associates is a digital storage analyst and business and technology consultant. He has over 39 years in the data storage industry with engineering and management positions at several companies. Coughlin Associates consults, publishes books and market and technology reports and puts on digital storage-oriented events. He is a regular storage and memory contributor for forbes.com and M&E organization websites. He is an IEEE Fellow, Past-President of IEEE-USA and is active with SNIA and SMPTE. For more information on Tom Coughlin and his publications and activities go to www.tomcoughlin.com.