WEKA

CASE STUDY

Contextual AI Builds Production-Ready Enterprise AI

Using Google Cloud and the WEKA Data Platform

About Contextual AI

Headquartered in Mountain View, Contextual AI's mission is to change how the world works through AI. The company provides a turnkey platform for building enterprise AI applications powered by its state-of-the-art RAG 2.0 technology.

The Business Challenge:

Drive Enterprise AI Adoption at Scale

Contextual AI is helping to solve the most pressing issues needed to enable enterprises to move AI into production, including model hallucination, currency, attribution, compliance, and data privacy.

Today, LLMs are being taught two things: continuity and coherency; they are great at providing responses that are clear and understandable, use correct grammar, and can maintain a coherent conversation. The problem is the current batch of models are not good at maintaining the most current state of information, and they lack context.

When AI models lack proper context for a query or task, the accuracy of the response drops quickly, and they extrapolate.

For instance, asking an AI model, "What's my fastest route to Paris?" without context, the model doesn't know if you mean Paris, France or Paris, Texas.

Extrapolation can lead to a model confidently making false but plausible-sounding statements— known as model hallucination. Existing models also include limited tools to highlight when a model response uses extrapolation and to show source material used to ground the response in facts.

Today, many developers leverage retrieval-augmented generation (RAG) to add external data to model responses in hopes of increasing the accuracy

contextual·ai

Challenges

- Long data load times
- Long model checkpoint write times

Solution

- WEKA® Data Platform on Google Cloud

Benefits

- 4x reduction in model checkpoint times
- Increased developer productivity
- 38% lower data cost per TB

of their large language models. However, a typical RAG system today uses a frozen off-the-shelf model for embeddings, a vector database for retrieval, and a black-box language model for generation, stitched together through prompting or an orchestration framework. This leads to a "Frankenstein's monster" of sorts. The solution is brittle, lacks domain-specific knowledge, requires extensive prompting and ongoing maintenance, and suffers from cascading errors. As a result, these Frankenstein RAG systems rarely pass the bar to make it into production.

Contextual AI's CEO and cofounder Douwe Kiela led the team that pioneered Retrieval Augmented Generation (RAG) at Facebook AI Research (FAIR) in 2020. Today, he is developing RAG 2.0 with the team at Contextual AI to address the inherent challenges with the original RAG system.

The company's approach focuses on two pillars: systems over models and specialization over AGI. Contextual Language Models (CLM) enable production-grade AI by optimizing the entire system end-to-end. With RAG 2.0, Contextual AI pre-trains, fine-tunes, and aligns all components as a single integrated system. As a result, customers can go from brittle generic chatbots to highly accurate and specialized AI applications, with improvements of over 4x compared to the baseline.

> "Training large-scale AI models in the cloud requires a modern data management solution that can deliver high GPU utilization and accelerate the wall clock time for model development."

The Technical Challenge

Contextual AI builds and runs its large language model on Google Cloud owing to the comprehensiveness of its compute infrastructure, which was purpose-built for AI/ML. Its AI model training environment consists of A3 VMs featuring NVIDIA H100 Tensor Core GPUs. Contextual relies on multiple datasets to train the next generation CLM across 7-70 Billion parameters.

AI workloads have significant computational requirements, making them time- and resource-intensive to train and serve them once training is complete. Contextual AI's original implementation was built in Google Cloud using its default storage offering, Google Filestore. However, they quickly ran into scale challenges and performance limitations, adding high costs and delaying its AI model development and training times.

In the fast-moving world of AI model training, speed often seems like the only thing that matters. Model providers ship new features and models weekly, so delays in training times can mean missing the first-to-market window. For most model providers, winning that race is all about bigger, faster GPU-accelerated infrastructure. However, with no end in sight to the mismatch between AI scaling and Moore's Law, AI model builders are building increasingly large GPU clusters, which in turn is driving up the price for the fastest GPUs and leading to their scarcity.

"Training large-scale AI models in the cloud requires a modern data management solution that can deliver high GPU utilization and accelerate the wall clock time for model development." says Amanpreet Singh, CTO & co-founder of Contextual AI.

Forward-thinking companies like Contextual AI realize the solution isn't as simple as getting your hands on more GPUs than your competition. It requires architecting an AI data infrastructure environment that can maximize their use. The company began scrutinizing its entire data stack–compute, memory, networks, and data storage–to eliminate training bottlenecks and feed its GPU clusters with more data faster than ever.

Model flops utilization (MFU) is a critical indicator for overall GPU efficiency and highlights growing bottlenecks from storage and networks in the current imbalance. 50% MFU is an impressive benchmark, but organizations end up paying twice as much and losing half the time since the GPUs sit idle.

Beyond faster networks, Contextual AI employs new approaches to maximize the use of local memory and eliminate latency, but these are challenging to do in the cloud, where customers have less control of placement.

Early on, fast data movement from storage to accelerator emerged as a key consideration for the Contextual AI team to drive faster AI model training times. Metadata handling, checkpointing, and data preprocessing were all critical points where the original storage architecture broke down.

> "With the WEKA Data Platform, we now have the robust data pipelines needed to power next-gen GPUs and build state-of-the-art generative AI solutions at scale. It works like magic to turn fast, ephemeral storage into persistent, affordable data.
>
> **AMANPREET SINGH,**
> CTO & CO-FOUNDER OF CONTEXTUAL AI

Python infamously utilizises a large amount of tiny files, causing it to be extremely slow to load files within Google FileStore. An "import torch" command would cause delay in load times of 10 to 20 seconds, creating a substantial impact on developer productivity over the time of a training epoch.

The lots of tiny files problem is one of those critical tasks that most legacy storage architectures aren't well equipped to handle. With training, the model rapidly iterates to find the right file, open, read, close, and move on. The native cloud file systems, while good for more sequential read/write operations, are not architected to meet the speed and scale required in AI model training.

Model checkpointing is another critical task that traditional storage systems are not equipped to handle. Checkpointing is essential to ensure resiliency during the training cycle, but it can cause AI models to stop training to complete the checkpoint. Really, really fast writes of a few very large model-weight files are key so that model training can continue. Long model checkpointing times meant the training would block for up to 5 minutes while the checkpoint was being written.

The Solution

The Contextual AI team didn't have to look far to find an alternative to their existing storage system. Working with their implementation partner, Accenture, the team developed an initial proof of concept (POC) to test the performance, scale, and cost savings of potential alternatives, including Google FileStore, DDN, IBM Spectrum Scale, local SSD, and the WEKA® Data Platform.

The POC environment consisted of 50 clients in parallel, each running a single threaded training epoch. WEKA was deployed across a basic 6-node cluster of C2-standard VMs in Google Cloud to drive data operations. The benchmarks for comparison consisted of the incumbent solution - Google Filestore - and IBM Spectrum Scale. After an initial test period, the Contextual team found that WEKA outperformed Google Filestore,

including 212% higher aggregate read bandwidth, 497% higher aggregate write bandwidth, 212% higher aggregate read IOPs, 282% higher aggregate write IOPs, and 70% lower average latency. While the alternative solutions had excellent performance in many areas, the overall strong performance for WEKA combined with a lower solution cost ultimately carried the day. Further, the team found the WEKA solution eliminated delays in model checkpoint completion, reducing model checkpoint times by 4x times, which could dramatically improve developer productivity.

Following the successful conclusion of its POC, Contextual AI now relies entirely on the WEKA Data Platform to manage all of its datasets for AI model training – a total of 100 TB. In the new environment, WEKA software runs on a 10-node cluster of GCE C2-std-16 VMs, which provide a high-performance data layer built on the NVMe devices attached to each VM, a total of 50 TB of flash capacity. The single WEKA namespace extends to an additional 50 TB of Google Object Storage, providing a scalable, affordable data lake to retain training data sets and the final production models.

Outcomes

With hundreds of model training epochs under its belt, the WEKA Data Platform is exceeding expectations for Contextual AI developers. It provides an enormous leap in data performance that directly correlates to increased developer productivity and faster model training times.

"With the WEKA Data Platform, we now have the robust data pipelines needed to power next-gen GPUs and build state-of-the-art generative AI solutions at scale. It works like magic to turn fast, ephemeral storage into persistent, affordable data." observed Singh.

The WEKA Data Platform combines low-cost object storage and high-performance flash-based storage in a single namespace and manages automated data tiering between them at a granular level. Its zero-copy, zero-tuning architecture supports every workload profile—handling metadata operations across millions of small files during model training and massive write performance during model checkpoint operations—in a single data platform. With a single copy of data, customers get the performance they need without over-provisioning storage resources. Contextual AI's model checkpoint times have been made 4x faster, cloud storage costs have dropped by 38%, and developers are more productive.

About the WEKA Data Platform

The WEKA® Data Platform removes the barriers to data-driven innovation through its advanced software architecture optimized to solve complex data challenges and streamline the data pipelines that fuel AI, ML, and other modern performance-intensive workloads.

The design philosophy behind the WEKA® Data Platform was to create a single architecture that runs on-premises or in the public cloud with the performance of all-flash arrays, the simplicity and feature set of network-attached storage (NAS), and the scalability and economics of the cloud. Whether on-premises, in the cloud, at the edge, or bursting between platforms, WEKA accelerates every step of the enterprise AI data pipeline – from data ingestion, cleansing, and modeling to training validation or inference.

Mind-bendingly fast. Seductively simple. Infinitely scalable. Sustainable. Spanning edge, core, hybrid, and cloud. The WEKA Data Platform helps to overcome complex data challenges and accelerate next-generation workloads to unleash your organization's imagination, creativity, and potential.

WKA396-01 07/24