AMD × CLOUDERA

FROM EDGE TO CORE: DEPLOYING CLOUDERA PLATFORM AGENTIC AI ON AMD EPYC™ CPU-POWERED INFRASTRUCTURE

As AI ingenuity expands, enterprises need safe, cost-effective ways to innovate at scale. AMD EPYC™ processors work with Cloudera's leading AI platform to power more AI workloads—with potential overall cost savings—enabling you to extend existing CPU-based infrastructure investments.

AMD EPYC IS READY FOR AI

With up to 192 cores per socket, AMD EPYC CPUs support massive parallelism, helping ensure the entire AI pipeline runs efficiently—from data preparation and ingestion to retrieval-augmented generation (RAG) inference. Readily available CPUs are ideal for working with small language models (SLMs) and medium language models (MLMs) with up to 20 billion parameters.

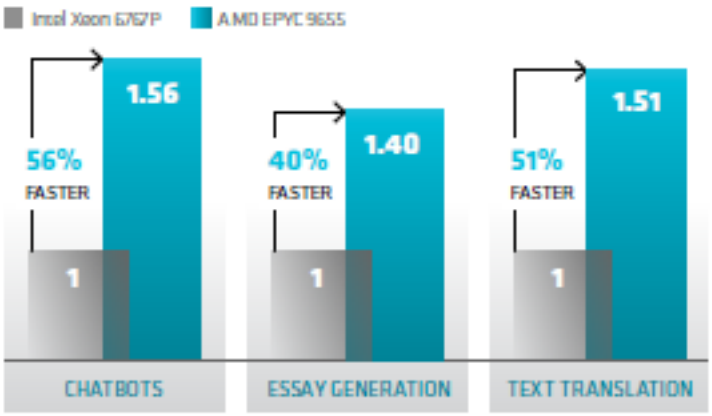
1.6x
SIMILARITY SEARCH PERFORMANCE[1]
FAISS with SIFT 1M dataset

1.9x
REGRESSION AND RANKING PERFORMANCE[2]
XGBoost on FP32

1.7x
FASTER END-TO-END AI PIPELINE[3]
Workload derived from TCP-x AI benchmark

Tests compare 2P 5th Gen AMD EPYC 9965 (192C) with 192 cores to 2P 6th Gen Intel Xeon 6980P (64C) with 128 cores. Results may vary.

OUTPUT TOKEN THROUGHPUT SPEED UP[4]



In tests using the Llama 3.1 8B model with ZenDNN and vLLM, single processor (1P) EPYC 9555 with 64 cores achieved significantly higher output token throughput than 1P Intel Xeon 6767P with 64 cores. SLA set at TPOT < 100 ms. AMD EPYC 9555 measured TTFT < 350 ms.

POWERFUL

Accelerated AI with Advanced Vector Extensions (AVX-512) and high memory bandwidth.

FLEXIBLE

Seamless processing at multiple stages of data-prep—feature extraction, tokenization, batching, vector search, and embedding.

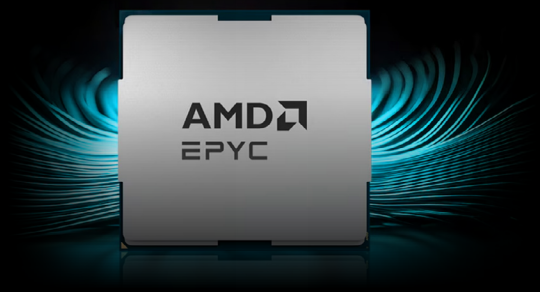
EFFICIENT

Leadership in energy efficiency across cloud, on-premises, and edge, enabling substantial cost savings.

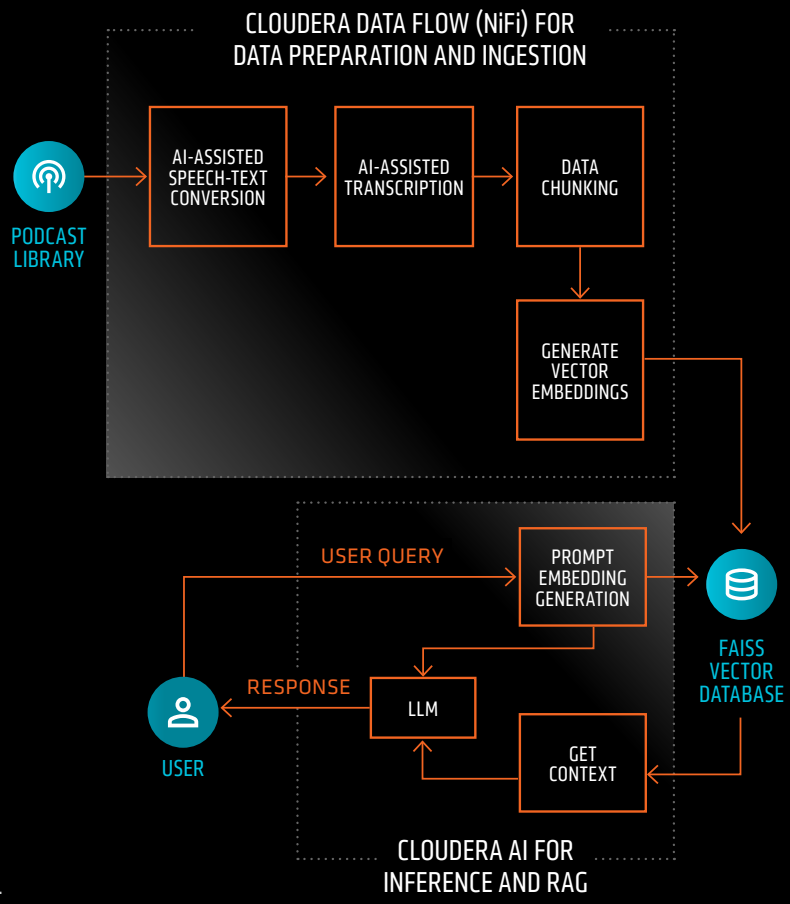
EXAMPLE USE CASE

To test performance, we used Cloudera Data Flow to orchestrate a popular use case—transcribing podcasts to text and building an AI enrichment pipeline for searching and chatting with the contents. The pipeline streams clean, real-time data into Cloudera AI Workbench for fast, accurate inference.

Cloudera AI workbench provides secure and collaborative AI development for accelerating model building, fine-tuning, and deployment of AI applications across hybrid or multi-cloud environments. In this scenario, users can search and chat with transcribed podcasts. The AI-enrichment pipeline uses Cloudera Data Flow to orchestrate the workflow, creating vector embeddings and storing them in a FAISS (Facebook AI Similarity Search) index. Cloudera AI is used for implementing RAG locally and inferencing.



Setup: 1P Dell PowerEdge 7615 with 1P EPYC 9354 32-core processor with 3.2 Ghz.



EXPLORE DELL POWEREDGE SERVERS

Whatever your data and AI needs, AMD and Dell provide options that are optimized for performance, energy consumption, and cost.

				
Model	Dell PowerEdge R6615/R6715	Dell PowerEdge R7615/R7715	Dell PowerEdge R6625/R6725	Dell PowerEdge R7625/R7725
Configuration	1U, 1-socket	2U, 1-socket	1U, 2-socket	2U, 2-socket
AMD EPYC processors	EPYC 9004 Series (R6615) EPYC 9005 Series (R6715)	EPYC 9004 Series (R7615) EPYC 9005 Series (R7715)	EPYC 9004 Series (R6625) EPYC 9005 Series (R6725))	EPYC 9004 Series (R7625) EPYC 9005 Series (R7725)
Memory capacity	12 DDR5-4800 up to 3 TB (R6615) 12 DDR5-5200 up to 3 TB (R6715)	12 DDR5-4800 up to 3 TB (R7615) 12 DDR5-5200 up to 3 TB (R7715)	24 DDR5-4800 up to 6 TB (R6625)) 24 DDR5-6000 up to 6 TB (R6725)	24 DDR5-4800 up to 6 TB (R7625) 24 DDR5-6000 up to 6 TB (R7725)
Front disk bay options (maximum)	10x 2.5", 4x 3.5", or 16x E3.S (R6615) 10x 2.5", 4x 3.5", or 20x E3.S (R6715)	24x 2.5", 12x 3.5", or 32x E3.S (R7615) 24x 2.5", 12x 3.5", or 40x E3.S (R7715)	10x 2.5", 4x 3.5", or 16x E3.S (R6625) 10x 2.5", 4x 3.5", or 20x E3.S (R6725)	24x 2.5", 12x 3.5", or 32x E3.S (R7625) 24x 2.5", 12x 3.5", or 40x E3.S (R7725)
Rear disk bay options (maximum)	2 (R6615) 2 (R6715)	4 (R7615) Not available (R7715)	2 (R6625) 2 (R6725)	4 (R7625) Not available (R7725)

RELATED LINKS

Cloudera > Dell PowerEdge Solutions Brief > AMD EPYC Processors > AMD EPYC Technical Briefs >

FOOTNOTES

[1] 9xx5-164. [2] 9xx5-162. [3] 9xx5-151. [4] 9xx5-255

AMD
together we advance_