IBM Watson Machine Learning Accelerator

Enterprise AI workloads at massive scale

Highlights

- Enables rapid AI deployment and includes popular deep learning frameworks
- Delivers faster insights by leveraging accelerated IBM Power Systems
- Supports large, more complex models with Large Model Support
- Simplifies deep learning training and inference across a cluster
- Shortens time to result with distributed hyperparameter optimization
- Allocates resources elastically
- Provides enterprise-grade multitenancy and dependability

Across many industries and professions, the data explosion is outstripping the human capacity to understand the hidden insights. AI can unlock the potential in all data—internal, external, structured, unstructured, voice, and visual—and make it work together. With enterprise-grade AI, organizations can make better operational decisions, understand customer wants and needs, communicate in real time, and optimize business processes—infused with the cognitive ability to understand, reason and learn.

IBM Watson Machine Learning Accelerator helps to make deep learning easier and faster for organizations by bringing together some of the most popular open source frameworks for deep learning, with development and management tools in a single installable package. Designed to simplify end-to-end deep learning, Watson Machine Learning Accelerator allows enterprises to spend less time on data preparation, implementation and integration, and more time training neural networks for results.

IBM Watson Machine Learning Accelerator includes the most popular deep learning frameworks in one installation:

- TensorFlow
- PyTorch
- Keras
- TensorRT

Distributed Deep Learning

Watson Machine Learning Accelerator easily distributes the training of models across many servers, leveraging many GPUs. Watson Machine Learning Accelerator brings intelligence about the structure and layout of the underlying cluster, which includes intelligence about the location of the cluster's different compute resources such as Graphical Processing Units (GPUs) and CPUs. As a result, customers can choose different distribution models based on their unique requirements. For example, elastic distributed training simplifies the distribution logic for the data scientist. The usage is simple: Define a maximum GPU count for training jobs and Watson Machine Learning Accelerator will schedule the jobs simultaneously on the existing cluster resources. GPU allocation for multiple jobs can grow and shrink dynamically, based on fair share or priority scheduling, without interruption. This is a unique IBM capability for deep learning. Distributed Deep Learning (DDL) is incorporated into the Deep Learning frameworks as an integrated binary, reducing complexity for clients as they bring in high-performance cluster capabilities.

Large Model Support

Organizations with AI workloads often face challenges with GPU memory limits (commonly 16GB). As models grow in complexity and data sets increase in size (high definition video vs web scale images), data scientists are forced to make tradeoffs to stay within the 16 GB memory limits of each GPU.

Watson Machine Learning Accelerator with Large Model Support (LMS), allows organizations to address larger data sets and more complex models. Enabled by NVIDIA® NVLink™ connection between the NVIDIA GPU and the POWER CPU (memory), the entire model and dataset can be loaded into system memory and cached down to the GPU for action. Users now are able to increase model sizes, data elements and batch or set sizes significantly, with the outcome of executing far larger models and expanding up to nearly one terabyte (TB) of system memory across 4 GPUs. Watson Machine Learning Accelerator with LMS opens the opportunity to address larger challenges and get much more work done within a single server, increasing organizational efficiency.

Watson Machine Learning Logo

Improved Data Ingest and Management

IBM Watson Machine Learning Accelerator includes an Apache Spark engine for data ingestion. This enables data to be manipulated in a resilient distributed dataset (RDD) across a large number of servers, using in-memory analytics. This approach provides all the power of Apache Spark, combined with full notebook and developer tool support, to ingest and manage data more effectively.

Elastic Distributed Inference

As a high performance, more reliable AI inference service, Watson Machine Learning Accelerator – Inference provides organizations with advanced scheduling logic to enable AI applications to scale inference calls on demand with near real-time decision making. Application integration uses an exposed API interface which supports batch, streaming, or interactive modes. Improved reliability and resiliency are provided for both the client connection and the service instances. With dynamic allocation of AI inference instances that are based on near real-time demand, Watson Machine Learning Accelerator – Inference offers high levels of resource utilization. Support for both CPU and GPU inferencing can further optimize the capabilities of the hardware platform.

Increased Infrastructure Efficiency

Watson Machine Learning Accelerator elastic resource allocation allows for the increase or decrease in resources while models are running, enabling clients to prioritize their workloads and make changes to allocations without having to restart training cycles. This elastic distributed scheduling is also extended to hyper parameter optimization. Distributed hyper parameter optimization shortens the time to accuracy by scaling out the hyper parameter tuning, allowing much higher number of iterations in a shorter time period. This creates a tremendous time-to-market performance value and increases the overall utilization of valuable hardware resources.

Hardware platform description and ordering information

Watson Machine Learning Accelerator is optimized for use with the following hardware configurations:

- IBM Power Systems servers (AC922 and IC922) with T4 or V100 GPUs
- x86-based servers with T4, P100 or V100 GPUs

Software Download
Direct Download
Watson Machine Learning Accelerator is distributed as a binary for Red Hat Enterprise Linux from IBM Entitled Software Support or IBM Passport Advantage.

Release Guide
A complete Release Guide with package list, prerequisites, deployment guide, and developer information is available at: www.ibm.com/support/knowledgecenter/SSFHA8

Why IBM?

For over a century, IBM has pioneered technologies and provided services that help companies manage and mine their valuable business data. And for 25 consecutive years, IBM has topped the annual list of US patent recipients—receiving a record-breaking 9,043 patents in 2017. In addition, IBM Power Systems are trusted by 78 percent of the Fortune 100. Further, every one of the top 10 banking firms have Power Systems, as do 80 percent of the top insurance and retail companies.

Next steps

→ Visit the website

For more information

To learn more about the Watson Machine Learning Accelerator, please contact your IBM representative or IBM Business Partner, or visit the following website: https://ibm.biz/EnterpriseAI

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: ibm.com/financing

This document contains information pertaining to the following IBM products which are trademarks and/or registered trademarks of IBM Corporation:
IBM®, IBM Power Systems™

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.