VMware Cloud™ on AWS for Machine Learning Workloads

AT A GLANCE

- Set up various machine learning environments and associated data infrastructure while enabling change as models are refined and evolved
- Eases incorporation of new ML technologies into the systems as they become available
- Use a tested high performing compute platform for Spark and other ML runtime platforms
- Run any version of code at any time
- Scale up compute power by re-sizing memory or virtual CPUs in VMs, or scale out using virtual machine cloning
- Virtual machine automatic restart and live migration of ML workloads between host servers
- Share compute resources across a variety of workloads using strict resource allocation policies

KEY BENEFITS

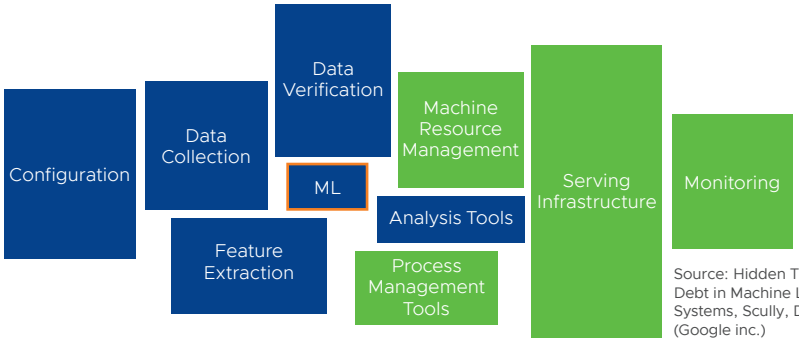
- Unifies the infrastructure on which everyone operates, ensuring standardization and reducing costs
- Minimizes changes and costs by running the learning phase and the production/ inference phase on the same infrastructure platform
- Maximizes ROI in cloud-based computing through higher utilization and greater availability
- Runs varied ML platforms, languages, and tool versions in parallel as virtual machines on a single common infrastructure with strong isolation
- Proven, familiar management environment for the administrator
- Ensures that virtual machine workloads can be brought back to life quickly using vSphere HA

Executive Summary

Machine learning platforms are rapidly developing, introducing new tools, features and versions that make it increasingly challenging for platform providers, data engineers, data scientists and systems administrators to keep pace with accelerating innovation. The requirement for multiple surrounding data engineering tools and infrastructure, as shown below, also contributes to the complexity. In the fast-changing machine learning environment, supporting compute platforms need to provide maximum flexibility. VMware Cloud on AWS fulfills the needs shown in green in the diagram below – and also provides a trusted compute platform for the remaining components.

Keep Pace with Innovation, Maximize Flexiblity

VMware enables data scientists and data engineers to choose different ML platforms, languages, tool versions and compute acceleration power at will – and run these all in parallel as virtual machines on a single common infrastructure with strong isolation. Data scientists and data engineers can rapidly scale-up compute power by re-sizing memory or virtual CPUs in their VMs, or scale-out using virtual machine cloning.



Source: Hidden Technical Debt in Machine Learning Systems, Scully, D., et al (Google inc.)

With capabilities such as virtual machine automatic restart and live migration, the VMware solution provides greater availability than is possible with physical environments. This enables data scientists and engineers to experience better SLA adherence, less downtime, more consistent performance, and easier repeatability of experiments. VMware also allows you to maximize your return on investment in IT resources through higher utilization and greater availability.

VMware Cloud on AWS is an on-demand service that enables customers to run machine learning environments across vSphere-based cloud environments with access to a broad range of AWS services. Powered by VMware Cloud Foundation, this service integrates vSphere®, vSAN™ and NSX® along with VMware vCenter® management, and is optimized to run on dedicated, elastic, bare-metal AWS infrastructure. ESXi hosts in VMware Cloud on AWS reside in AWS availability zones (AZ) and are protected by vSphere HA.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com