GIGAOM RESEARCH

The industry leader in emerging technology research

SUPPORTING DATA STORAGE ARCHITECTURES FOR ML/AI

MARKET LANDSCAPE REPORT

AUTHOR: Enrico Signoretti - GigaOm Analyst

AUTHORED BY GIGAOM ANALYST, ENRICO SIGNORETTI

INTRODUCTION

With the growing interest in machine learning (ML) and artificial intelligence (AI) in enterprise organizations, the market is quickly moving from projects and infrastructures designed for research and development to product and turn-key solutions designed by vendors to answer quickly to new business requests. Enterprises understand the potential of ML and AI to boost their market competitivity or to improve several internal processes that are now time demanding. At the same time, software tools are getting better and more user-friendly making it easier to find the necessary skills to build new applications or reuse existing models for more use cases.

The market is maturing quickly as well. High-performance computing (HPC) vendors are now joined by other manufacturers, usually focused on enterprise workloads, and startups in a race to offer the best solutions to end users that usually purchase smaller storage systems, or cloud storage, and are fairly new to HPC computing. In fact, even though some infrastructure design concepts are similar to what is required for big data analytics, the specific nature of ML/AI algorithms, and GPU-based computing require more attention to throughputs and $/GB, especially because of the sheer amount of data involved in most of the projects.

Depending on several factors, including an organization's strategy, size, security needs, compliance, cost control, flexibility, and so on, the infrastructure could be entirely on-premises, in the public cloud or a combination of both. In fact, the most flexible solutions are designed to run in all these scenarios, giving the end user ample freedom of choice. Long term and large capacity projects, run by skilled teams, are more likely to be developed on-premises while smaller teams typically use the public cloud for its flexibility and less demanding projects.

The efficiency of infrastructures for ML/AI workloads is key to reducing the time it takes to see results. Except for the initial data collection, many parts of the workload are repeated over time, with latency and throughput that are crucial for the entire workflow. Latency is important to handle metadata quickly, while throughput must be as high as possible to ensure the system GPUs are always fed at their maximum capacity. In fact, a single modern GPU is a very expensive component, capable of crunching data up to 6GB/s and more, and each single compute node can have a few of them installed. In other words, CPU storage vicinity is another important aspect of these architectures and this is why NVMe-based flash devices are usually selected for their characteristics of parallelization and performance. On the other end of the spectrum, there is a huge amount of capacity needed for storing all the data sets for the training of the neural network. In this case, scale-out object stores are usually preferred because of their scalability characteristics, rich metadata, and competitive cost.

In this report, we will discuss the most recent architecture designs and innovative solutions for storage infrastructures deployed on-premises, cloud, and hybrid fashions, aimed at supporting ML/AI workloads for enterprise organizations of all sizes. We will analyze two-tier and single-system architectures, their respective advantages and disadvantages and how they can be integrated with the rest of the on-premises infrastructure or the cloud initiatives already in place.

REPORT TOPICS

- Enterprise organizations are aware of the strategic value of ML/AI for their business and are increasing investments in this area.
- End users are looking for solutions that are easy to implement and use, ready to go, and with a quick ROI.
- Depending on the project size and other requirements, solutions can be on-premises, in the cloud, or hybrid. The goal is to get efficiency while keeping costs at bay.
- Most of the solutions available in the market are based on two-tier architectures with a flash-based parallel scale-out file system for active data processing and object storage for capacity and long term data retention.

CONSIDERATIONS FOR ADOPTING SOLUTION
(IN QUESTION FORM)

- Are you evaluating AI/ML techniques for your organization?
- Do you want to keep control of the entire AI/ML workflow?
- Will you develop your AI/ML project on the cloud, on-premises, or both?
- Do you already have an estimation of the data set size and storage capacity required?

ABOUT

ANALYST ENRICO SIGNORETTI

Enrico has **25+ years of industry experience** in technical product strategy and management roles. He has advised mid-market and large enterprises across numerous industries and software companies ranging from small ISVs to large providers.

Enrico is an **internationally renowned visionary author**, blogger, and speaker on the topic of data storage. He has tracked the changes in the storage industry as a Gigaom Research Analyst, Independent Analyst and contributor to the Register.

@esignoretti

INTERESTED IN GIGAOM REPORTS?

To purchase this report, or to explore opportunities to participate in future GigaOm reports, Email a GigaOm Business Development Representative.

GIGAOM

Find us:
gigaom.com

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization. To apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.