By Merv Adrian
Founder and Principal Analyst, IT Market Strategy

The VAST Data Platform Redraws the Data Infrastructure Landscape

Unified Storage, Database, and Compute Engine Services for the Future of AI

Reimagining Systems Architectures

The history of data infrastructure architecture has been one of continuous evolution, but the past few years have witnessed a dramatic discontinuity brought about by technology advances, revolutionary software approaches driven by open standards, and the move to cloud and hybrid architectures. These factors have made it possible to fundamentally rethink how data is created and managed.

VAST Data has emerged as a leader in a remarkably short time by finding the synergies across emerging hardware and software technologies, coupled with an innovative approach to ecosystem-based delivery that has changed their cost model, and those of their customers, dramatically. Their original offering, dubbed Universal Storage, rethought the fragmented landscape and pointed the way towards a more effective and consolidated storage architecture that promises to eliminate "tradeoffs" – or more precisely, compromises – that have held development and deployment back.

Now VAST Data is moving to the next stage with the VAST Data Platform, moving up the stack to tackle a growing set of new data infrastructure challenges beyond storage.

Scope Shifts: We Need It All

The marketplace has spawned more data and new types of data at an increasing rate, and use cases have expanded their needs. Instead of dedicated collections of data, increasingly use cases need more – often unplanned – access. This wider usage fuels expansive requirements for real time analytics, sophisticated semantics across disparate sources, and revolutionary changes in Artificial Intelligence that demand access to "everything, all the time."

In recent research, IDC noted that 51% of respondents anticipated data growth between 20% and 49% in 2022, while 31% suggested higher rates than that: 50% to 99%[1]. If anything, these expectations may have been low. Proliferating AI systems have fueled a massive appetite for additional data usage as evidence mounts that large language models (LLMs) will be error-prone – the term of art is "hallucinations" – unless they are fed more data. Much of that information is in unstructured form, and much of it will initially be deployed as copies of data already stored.

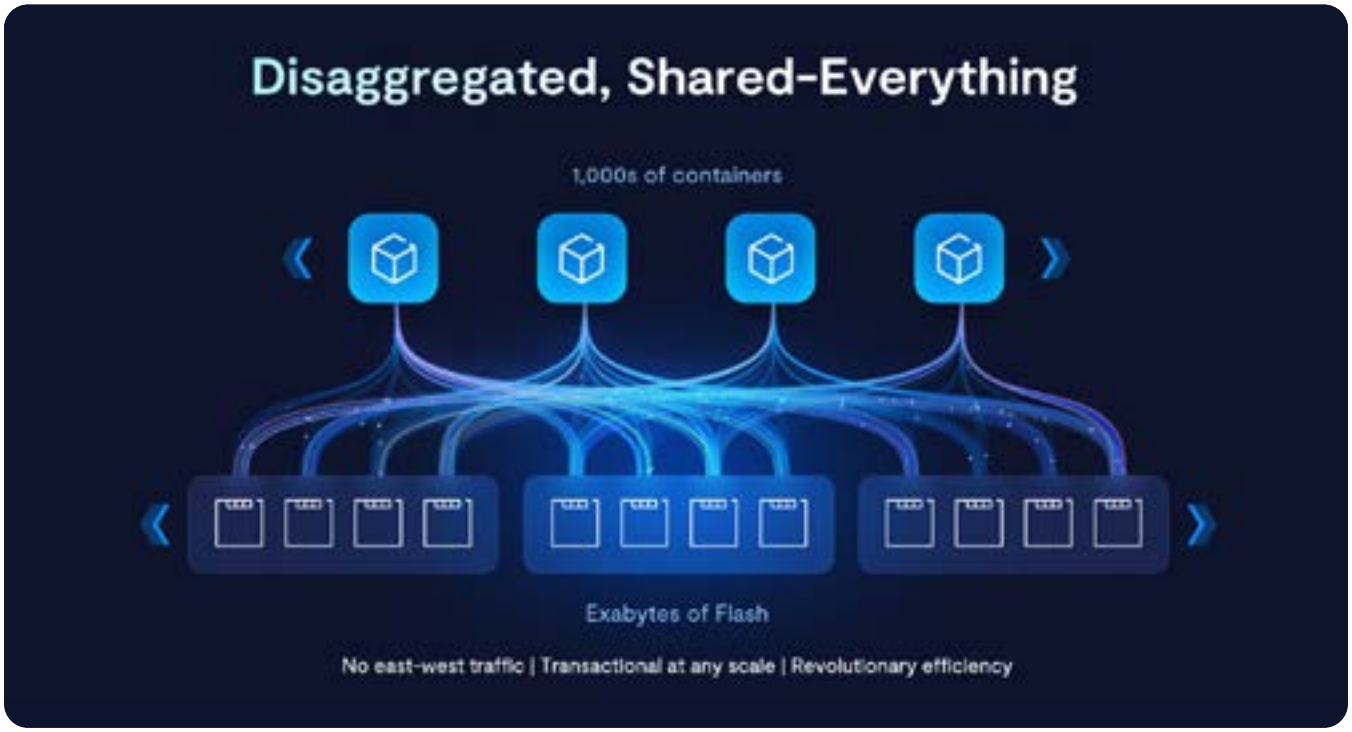
Equally significant, this flood of added data will need to be backed up, timestamped, and available quickly to respond to unexpected new requirements, geographic expansion, and deployment alternatives. Data that has been moved to the cloud may suddenly be required for on-premises requirements driven not by technology but by compliance, or by application stacks that reside on "other" clouds.

The promise of the cloud to separate compute and storage had a profound implication if carried to its logical conclusion: **Information ought to be available at the lowest level of detail to any compute resource, and importantly, to many of them, all the time, simultaneously.** In practice, that separation turned out to create new tradeoffs that continued to delay the realization of that vision.

The VAST Data Platform

Years of software efforts at data engineering have led to complex architectures that shred, duplicate, reformat, tier and place data in multiple locations. These are fragile and labor-intensive processes that lead to performance challenges, fragility and high costs. VAST Data's architectural innovations introduce a new level of automation that unifies performance optimization with usage optimization – with substantially lowered requirements for human intervention.

VAST Data's fundamental innovation at the storage layer, now in use in enormous deployments worldwide, is described as disaggregated shared-everything (DASE). DASE is an all new, distributed system architecture that eliminates tradeoffs between performance and capacity, between performance and global consistency, and between data-driven and event-driven architectures.



A key concept for readers accustomed to existing strategies for dealing with large data volumes is the idea of "no east-west traffic." The replication strategies of distributed DBMS, and the familiar scatter-gather model of MapReduce bounce off this wall – it is a performance killer, a source of complexity, and a barrier to widespread use of distributed data. VAST's model eliminates the need for "shuffle steps" and long merge journeys, even as its reliance on flash speeds up the retrieval times themselves.

When costs drive architectural decisions, the result is often a tradeoff between spending and performance. The DASE architecture helps attack this boundary with architectural simplification: as a single tier, it can consistently deliver the required performance without adding complex configuration and management to the equation. And with it as a foundation, additional layers of the stack can be unlocked, with offerings that rely on its innovations.

This fully realized stack is the **VAST Data Platform** that contains four core components: the **VAST DataStore, VAST Database, VAST DataSpace and VAST DataEngine.**

The VAST DataStore

To begin the journey, it's necessary to break down the barrier between databases and file systems and unify them to enable more sophisticated use of data that hitherto was in one place but not the other, on either side of the "structured vs unstructured" boundary.

This begins with the VAST DataStore, which provides the multi-protocol global namespace, the metadata index/catalog and the physical services that underlie the architecture: replication, snapshots, load balancing, data protection, reduction of volume via multiple techniques leading to a highly compressed and rapidly accessible data asset.

All of the data "inside" the VAST DataStore is encrypted in flight and at rest, authenticated and controlled by external key management. The system provides load balancing, manages remote and local clones, and uses standard protocols such as NFS, S3 and resilient SMB.

The metadata in the VAST DataStore extends the model beyond that provided by the file systems that underlie today's data deployments. Its rich capabilities enable the layers above to perform their duties in entirely new ways and delegate some of them "to the storage layer" in ways that liberate them to reduce their code paths and dramatically improve their reach and performance.

The VAST DataBase

The VAST DataBase begins by processing incoming data in a row buffer, where transactions can be acted on instantaneously, and then migrates them to columnar storage, organized in small units (32K apiece) that include metadata to aid rapid retrieval, sorting and filtering. The design adds the table directly as an element to the file system. Early tests have shown it outperforming recent formats like Parquet files inside S3 using the Iceberg table format, another columnar data layout commonly used for tabular data.

The VAST DataBase adds another access protocol – SQL – to the vocabulary of the DataStore, permitting operations that span the "structured and unstructured" categories. With metadata that can correlate hitherto separated domains such as satellite imagery, genomes, video feeds, trade data, the VAST DataBase enables richer understanding both for relatively traditional developer-led applications and for the insatiable need of AI to ingest everything that might be relevant to ensure its accuracy.

This is no small point. Any observer of the current frenzy over generative AI and LLMs will have encountered the term "hallucination" that identifies the unfortunate tendency of such systems to be incorrect if not given the data required to make their conclusions accurate. Retrieval Assisted Generation, using models that combine pre-trained parametric and non-parametric memory for language generation, is one path to improving this situation. (Experienced readers may be forgiven for thinking "isn't this just the old saw 'garbage in, garbage out' in new words?" Yes, but it's more about remediation than identifying the problem.)

How to make the most of this direction? Make all the data available, all the time.

The VAST DataSpace

Another element is needed to complete this picture: global deployment and global access that breaks the boundaries between private and on-premises deployment and multiple clouds. To traverse today's landscape in an often unavoidably multi-deployment world, enterprises must deal with multiple APIs applied differently in different places, accept latencies these traversals introduce, and contend with differences at the edge: high speeds from multiple points and sudden bursts that overwhelm systems designed to different assumptions.

Lock management is the barrier in most systems – even with relatively local scope it introduces complexities for the preservation of consistency, and across multiple sites it requires enormous amounts of management to achieve even adequate performance, often by relaxing business requirements to get there.

The VAST DataSpace introduces fine-grained lease management for files, objects and tables to this picture. Each of these can be leased by location that needs them while providing consistent access to data from all locations. The cloud ecosystem VAST Data has organized with AWS, Azure and Google Cloud is available now.

The VAST Data Engine

One final component remains – bringing immediacy to the system by having it encompass computation from an event-driven perspective. The premise of databases for decades has been that they exist to store data at rest, to provide a platform to add, maintain, change and report on that data, and enforce policies to ensure accuracy, quality and trust. In effect, except for a moderate subset, they existed downstream from the actual activities that created the data in the first place.

As a result, an entire category of software lives "upstream" – sensor networks, social platforms, and myriad activities that 'just use file systems." These are often append-only, and associated with highly specialized applications that are used to record them, encode them, search and retrieve them to be handed to other specialized uses.

Unifying these two different views becomes possible when functions are viewed as elements to be stored, and events associated with files can be used to cause them to be run. The VAST Data Engine uses conventional terms for these two ideas: functions and triggers. But the application of these elements now has universal scope, across anything that can be kept "in the VAST Data Platform" – which is to say, "any data."

The idea is not new – "content as code" is a familiar idea in computing circles. It offers the promise that workflows can be self-actuating, and that the reuse of functions can be facilitated by keeping them closely associated with the data. This promise is realized in the VAST DataEngine, due to its ability to run new data automatically to specific functions as it arrives, based on metadata the system manages.

Moreover, changes caused by the processing of the data may themselves trigger additional activity. Again, this is not a new idea – inside a DBMS the update or addition of a data element might trigger the update of an index, the adjustment of an optimization algorithm due to changes in statistics that activity created, and so on. What is different here is the breadth and universality of the functional reach.

Enabling the New Architectures

Why does all of this matter to AI applications? Because models cannot be static. They need to evolve – this is the heart of learning systems.

Today, a great deal of effort is expended on "training" models – but that is only the beginning. Those models can go out of date and must be maintained. Who will do it? Why not the system itself – this is a true model of a learning system. And its scope is a critical requirement – we can teach it not just to react to "feeds" we have already set up, but to go find data it needs to resolve uncertainties. Conclusions – the output of logical functions and inferences – will be stored and themselves become source material. Such capabilities could be self-actuating – creating and executing new iterations of the model over data that was already processed to leverage new insights about historical data and predict the future better.

Many early AI-based projects have relied upon one-time imports of data from relatively "dark" sources, ones that have hitherto been rarely exploited because they are different in format, stored on other platforms, or external to the organization that uses them. The opportunity to truly operate on "any data" has been a theme of data analysts and architects for years. But until now it has been infeasible at scale, across platform boundaries and data types unless enormous effort was expended, one project at a time. And the one-time imports aged quickly. At best, complex new change data capture strategies needed to be created to meet the requirements for current data – or the entire complex build needed to be replicated and another large data source copied.

The VAST Data Platform changes all of this. Current data, growing explosively in real time, across platforms, being used without unnecessary, fragile, costly replication strategies will redefine the art of the possible. A new era of fully exploited data resources will unleash the transformational power of emerging technologies like knowledge graphs, comprehensive enterprise search and generative AI. Moreover, it will give such technologies the scope to deliver on the promise of augmented applications, analysis and operations, with an impact as profound on everyday life and business processes as the industrial revolution – and more far-reaching.

Existing technology platforms have begun to approach the end of their design paradigms' utility, and organizations worldwide are ready to modernize all the way down the stack. VAST Data will enable them to rethink and break down barriers that were long assumed to be unchangeable.