White Paper

Digital Transformation Drives New Fast Data and Big Data Storage Platform Requirements

Sponsored by: Western Digital

Eric Burgener
February 2018

IDC OPINION

Enterprise information technology (IT) is in the midst of a full-blown digital transformation. Legacy applications like relational databases, collaboration platforms, and other latency-sensitive workloads must continue to be supported while CIOs bring on next-generation applications. These next-generation applications include mobile computing, social media, big data and analytics, and cloud-based workloads. All these structured data workloads are being consolidated onto virtual infrastructures, and an increasingly internet-literate customer base is driving the demand for ever faster response and agility. Hard disk drive (HDD) technologies cannot cost effectively handle these random, varied, and latency-sensitive workloads, a fact that within just the past several years has driven the use of all-flash arrays (AFAs) to dominate these new performance-sensitive "fast data" environments. At the same time, the increasing use of big data and analytics is driving transformational changes in unstructured data environments that are changing the structural storage requirements for these less latency-sensitive workloads. IDC refers to this emerging tier of big data-driven workloads as "big data" environments.

As enterprises evolve their IT infrastructures to better accommodate these workload types that will dominate commercial computing, vendors will be increasingly introducing storage solutions that conform to the new definitions of "fast data" and "big data" platforms. These definitions share some capabilities with traditional primary and secondary storage platforms but go beyond them in important ways. These newer platforms will need to leverage newer technologies, such as flash media options, software-defined storage designs, self-driving storage paradigms, and scale-out architectures, while delivering on a "five-nines plus" availability metric. Data services like flash-optimized RAID and erasure coding options, inline data reduction, space-efficient redirect on write snapshots, quality-of-service (QoS) features, and encryption – all of which can be applied selectively at the application level – will be baseline requirements for fast data platforms, while massive scalability, low cost per gigabyte (GB), and a different set of data services, including capabilities like versioning, audit trails, and write once, read many (WORM), will be must-have features for big data platforms. Storage tiering options will be required on both types of platforms. As customers evolve their IT infrastructures for the future, a good understanding of the developing fast data and big data definitions will help customers inform better purchase decisions.

IN THIS WHITE PAPER

As enterprises undergo digital transformation, they are moving toward a future that will have production business applications that fit into two broad classes: fast data and big data. Fast data workloads include mission-critical primary applications requiring consistently low latencies, while big data workloads are composed of applications that tend to be less latency sensitive and leverage much larger data set sizes. This white paper discusses the storage infrastructure requirements for fast data and big data platforms as well as the storage requirements around them that businesses will need to meet going forward.

SITUATION OVERVIEW

Today, having the right IT infrastructure is more important than ever in driving business success. Enterprises must accommodate the need for a primary storage infrastructure for latency-sensitive primary workloads that directly drive revenue and other critical metrics while maintaining what has traditionally been considered a secondary storage infrastructure. This latter infrastructure hosted less latency-sensitive workloads like backup, disaster recovery, and archive that were targeted at data protection and regulatory/compliance requirements and typically needed to support multi-petabyte (PB) data sets over longer periods of time than primary workloads. Business analytics applications have been a staple of enterprise IT infrastructure since the 1970s and represented somewhat of a middle ground between primary and secondary storage since they ran on top of large data sets considered critical to business success but did not generally have to meet low-latency requirements.

Over the past five years, the IT industry has seen a major change in the commercial analytics arena. With the pace of technology evolution, companies can now cost effectively collect massive amounts of data on their products, processes, and customers, and the Internet of Things (IoT) is making this data readily available for analysis purposes. Businesses are actively pursuing big data and analytics as a way to better understand customer and market requirements, deliver improved service to their clients, and inform better business planning. Small sampling sizes in the past have imposed limitations on the accuracy of commercial analytics, but big data increases the validity of analyses since it effectively provides access to nearly unlimited data points over periods of potentially many years. Technology advances in the compute and storage arenas have opened up opportunities to leverage these massive data sets in a more real-time manner, allowing some companies to build competitive advantage around their ability to rapidly respond to changing market conditions and customer preferences. Big data is clearly the future.

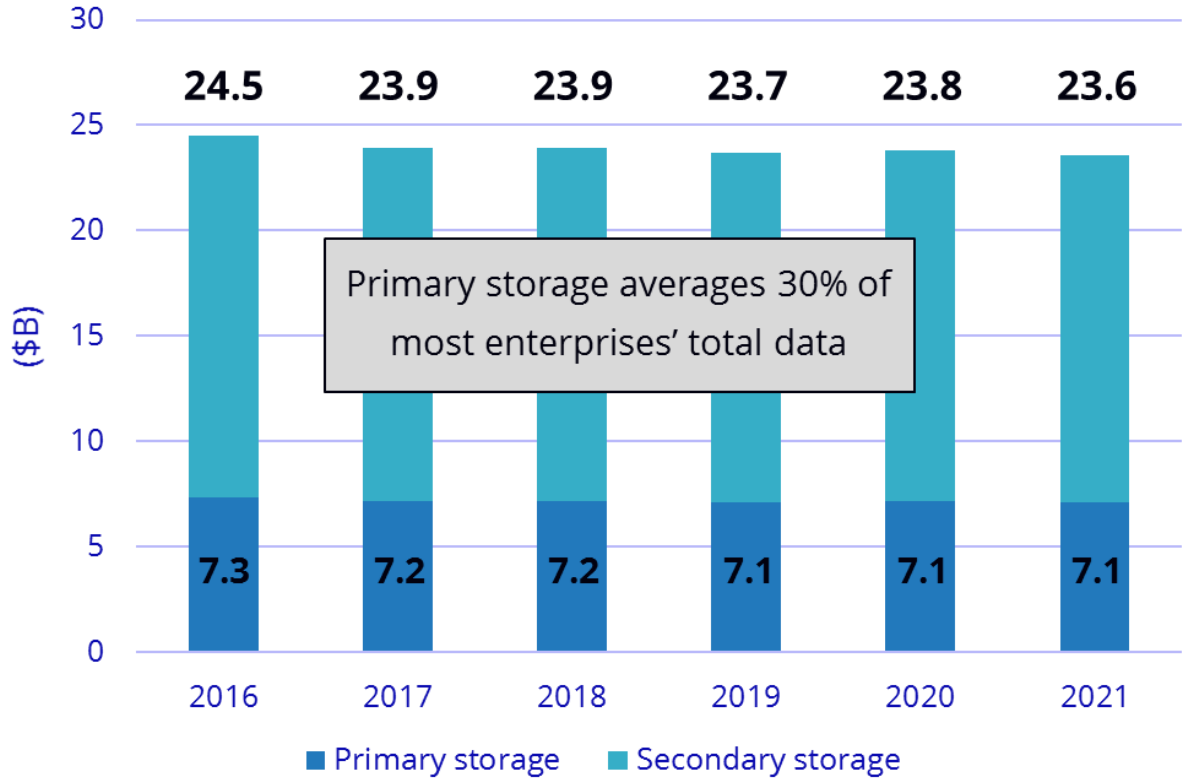
Newer workloads for mobile computing, social media, big data and analytics, and cloud are driving a significant need for IT agility in the enterprise. The hardware-defined, more static IT configurations of the past are giving way to a more software-defined and much more agile future that makes businesses much more responsive to customer and market demands. The software-defined datacenter, built around either commercial or open source virtualization infrastructures, offers clear advantages for today's dynamic business environment. Virtually all IT organizations run virtual infrastructure, have undertaken storage consolidation projects to move legacy workloads to the infrastructure, and deploy a "virtual first" strategy for most new application deployments.

Today, most servers running both primary and secondary applications are virtualized. The storage platforms on which the primary and secondary data sets are stored, however, are often separate. Historically, latency-sensitive primary workloads have been hosted on block-based storage platforms

that deliver low latency for transactional application environments, support transparent recovery from failures, offer very high availability and reliability, and have generally supported capacities in the hundreds of terabytes (TB) to low petabytes. Secondary storage platforms, on the other hand, generally supported file- and/or object-based data sets and were designed chiefly to support scalability into the tens of petabytes and beyond with much lower cost per gigabyte. Software functionality also varied across the platforms, geared specifically to the types of workloads they supported. Both markets (primary and secondary storage) are growing, multibillion-dollar markets, as shown in Figure 1.

FIGURE 1

External Enterprise Storage Revenue by Primary and Secondary Storage, 2016–2021



Source: IDC, 2018

Primary "fast" data makes up roughly 30% of overall external enterprise storage capacity, with secondary "big" data making up the rest.

With the increase in data overall that enterprises want to store and actively leverage, increasing administrative costs are also a concern. Automation and orchestration tools provide an excellent way to keep administrative costs low while improving the reliability of operations. Many newer software-defined systems employ a "self-driving storage" concept. In these systems, storage operations are driven by business policies that use defaults, templates, and dynamically adjusting responses to

workload variances, failures, and other events that in the past have required manual intervention. Once an objective has been identified, the system automatically takes the necessary actions to meet it. These types of "self-driving storage" designs are the future, going a long way toward increasing the administrative span of control (i.e., how much storage a single administrator can manage) in today's data-driven environments.

As the digital transformation continues to unfold, the primary and secondary storage platform definitions will need to evolve as big data and analytics continues to penetrate mainstream commercial computing environments. As these two workload classes (fast data and big data) develop over time, enterprises will need to ensure they have the right compute, storage, and network resources available in their IT infrastructure to effectively service them. The traditional definitions of "primary storage" and "secondary storage" will evolve over time into the "fast data" and "big data" definitions. We'll now turn to the requirements of fast data and big data workloads.

Fast Data Workload Considerations

One of the key impacts the internet has had is in driving customer expectations for real-time response. Both mobile computing and social media support a much more densely connected world and have helped instill in users an expectation of immediate gratification. This new climate has significant implications for business agility in ways that impact revenue generation, the pace of expected innovation, responsiveness to service issues, and overall customer satisfaction with products and services. Within the typical enterprise application portfolio is a set of workloads that require uniquely high performance, and this has always been true. In today's more time-sensitive environment, there are real business reasons why certain applications need even, consistent sub-millisecond storage latencies.

With the dense workload consolidation that has gone on in the evolution to virtual computing, I/O profiles have become much more random as well, and it is clear that enterprises cannot cost effectively meet these requirements with HDDs. Noisy neighbor problems (i.e., when I/O spikes in one application unexpectedly impact the performance of other applications resident on the same virtual infrastructure) are a major concern with these types of latency-sensitive workloads, potentially impacting a company's ability to meet service-level objectives (SLOs). The industry response has been the rapid migration to the use of all-flash configurations in primary storage platforms of all types, as well as the rise of self-driving storage approaches that dynamically manage systems to meet defined performance levels.

Different storage media types should be available and configurable to meet different performance requirements. While most primary storage platforms over the past several decades have used SCSI-based HDDs, over the past several years there has been significant innovation in storage media. Solid state disks (SSDs) with SCSI interfaces are dominating media types in primary storage platforms today, delivering at least an order of magnitude better performance than SCSI-based HDDs. The SCSI protocol was built specifically for HDDs, however, and does not deliver all the performance that flash media is capable of. The NVMe protocol was developed specifically (and only) for flash media and operates much more efficiently than SCSI to deliver noticeably better latencies and throughput for most workloads than SCSI-based SSDs can deliver. NVMe-based devices are, however, more expensive than SCSI-based ones today.

Storage class memory (SCM) is another new technology that will start to appear for production use on enterprise storage platforms in 2018. In terms of performance and cost, SCM takes a middle ground between DRAM and NVMe devices. While single-tier AFAs have been driving most of the primary

storage revenue over the past couple of years, during 2018 we will see the release of more tiered storage architectures that leverage a mix of different media types like SCM, NVMe, and SCSI. To cost effectively accommodate the need for different performance levels, IT infrastructure will need to offer all three types of media as disaggregated resources, which can be allocated as necessary when defining new virtual servers.

Most primary storage platforms today already include a full panoply of enterprise-class data services. Algorithms that leverage data redundancy, like RAID and erasure coding, provide a first line of defense against device failures and include implementations that offer configurability in terms of how the data is laid out and allow administrators to "tune" their data protection to meet specific requirements, like lower latency, improved resiliency, and lower cost. Storage efficiency features like compression, deduplication, pattern recognition, write minimization, and space-efficient snapshots and clones increase storage density, thereby lowering the effective cost per gigabyte to store data, and can also lead to faster, more responsive operations, improved media endurance and, in some cases, better performance.

Encryption can help meet regulatory and compliance requirements and can be implemented either in software, with hardware assist in controllers, or with self-encrypting devices — each of which offers its own pros and cons. Quality-of-service controls allow certain workloads to be prioritized over others to meet SLOs when they exist and offer the ability to set other parameters affecting performance such as latency, throughput floors and ceilings, and performance classes that make provisioning new storage easier. Replication is another key data service in latency-sensitive mission-critical environments, providing options for data protection and disaster recovery as well as data distribution, which can help lower access latencies for geo-distributed workloads. On virtual infrastructure, customers will generally be configuring virtual servers that are dedicated to a particular application and will need the ability to apply data services selectively at the application level.

With the march of technology, IT infrastructure density is increasing. Infrastructure density measures how much "work" can be done in a given footprint (i.e., how much throughput is required and storage capacity is located in 1U of rack space). As processors get more powerful, storage gets denser and network throughput increases, allowing a single server to perform more work. This drives IT costs lower but also raises concerns about "failure domains." As a single server can drive much more work, it also increases the impact of that server failing. One of the key concerns for fast data workloads is that they operate at predictable, defined levels of performance all the time. They must be able to nondisruptively work through failures, maintenance tasks, system expansion (to accommodate rapid data growth), and other events that in the past may have caused service disruptions. This means that the fast data infrastructure must not only be very high performance but also very highly available.

In the future, storage architectures are likely to depend more on distributed designs that spread the workload across a high number of components. These types of designs help not only deliver more consistent performance with varying workloads but speed rebuilds from device failures because recovery benefits from high degrees of parallelism. Self-healing architectures that can quickly return to full-protected-mode operation (by rebalancing the workload across all remaining components) will also become more popular.

Fast data workloads include legacy applications, like transactional databases that directly drive a key business metric such as revenue generation, as well as newer, more latency-sensitive next-generation applications in the mobile computing and social media areas. Internet applications that interact with customers in real time demand consistent sub-millisecond response times under load to make

websites and other online activities "sticky." Instant messaging and some other collaboration platforms may also be considered fast data workloads. The key feature of the fast data class is that it must meet stringent performance requirements even in the face of densely consolidated and widely varying workloads, and it must do this even when data services such as data reduction, snapshots, and encryption are in use.

Big Data Workload Considerations

Big data environments have historically been associated with large, unstructured data sets (file and/or object), with much more of a focus on bandwidth and throughput than latency, and massive scalability. These systems were typically not considered mission critical, so while they offered data protection options to maintain data integrity, they typically were not built to provide the same kind of rapid recovery that fast data workloads require. This is still true, but there is a growing class of big data workloads that are both real time and considered mission critical, and IDC sees these next-generation big data workloads becoming more pervasive over time. IDC expects that by 2020, 70% of the Fortune 2000 will have at least one real-time big data and analytics workload that they consider to be mission critical, with a smaller percentage depending on multiple real-time big data and analytics applications that are mission critical.

Because of the high throughput and bandwidth requirements against very large data sets, many big data platforms use scale-out architectures. Throughput and bandwidth can be easily scaled by adding more compute and/or storage resources, and distributed software designs balance workloads across them, driving higher performance through the use of massive parallelism. Disaggregated scale-out storage infrastructures provide significant flexibility for administrators in achieving the optimal balance of compute and storage for any given set of workloads, driving cost efficiencies that can make an appreciable difference in large-scale big data and analytics environments.

Different storage media types should be available and configurable to meet different performance, capacity, and cost requirements. On a cost-per-gigabyte basis, HDDs are still less expensive than SSDs, but flash media prices are continuing to drop. For big data workloads that care more about cost than they do performance, 7,200rpm SATA HDDs (at $0.02 per gigabyte) are hard to beat, but given the multi-petabyte data sets that are becoming common in big data environments, their limited capacity means that you need a lot of them. One of the benefits of flash media, particularly with the multidimensional 3D NAND packaging, is its extremely high density. While today's largest HDDs are 12TB in size, IDC expects to see 64TB commodity off-the-shelf SSDs available for production use in 2018. When scaling into the tens of petabytes and beyond, these much larger device sizes drop flash cost per gigabyte even lower, require much less energy and floor space, and result in more reliable infrastructure because there are fewer devices required to meet capacity requirements. Larger device sizes, however, could introduce recovery time concerns, depending on vendor implementation. When configuring storage platforms to support massive scalability, customers will want a range of both HDD and SSD options. The availability of SATA, SAS, and NVMe interfaces on the large-capacity SSDs will also provide more ability to tune these environments for more performance for real-time I/O, data movement, or device rebuilds (on failure) if and when it is required.

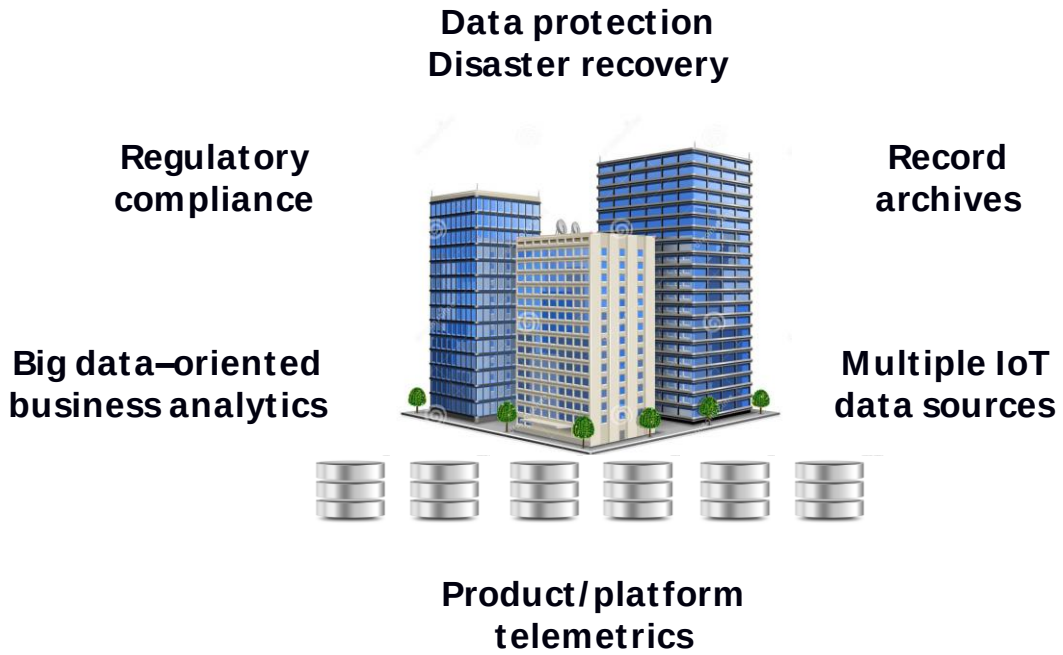
Massively scalable platforms will also benefit from the availability of other media types. Cloud storage should be accessible as an optional tier because, for some data sets, it provides less expensive options for compute-intensive operations and data retention. When working with large data sets that may need to be moved to the cloud and back, the high bandwidth of NVMe, the space-saving benefits of data reduction and delta differential technology, and the massive parallelization available with

today's compute, storage, and networking platforms help enable rapid data mobility in geo-distributed environments. Many storage platforms today support an auto-tiering capability that can migrate data to different tiers when specific parameters (e.g., low frequency of access) have been met. For long-term data retention with data sets that must be retained over decades and are not likely to be accessed very frequently or with any sense of immediacy (electronic medical records, court and DMV records, etc.) offline media options such as tape may be attractive as well due to their extremely low cost. Any data sets that are required to support real-time analytics, however, must remain on online media.

There are a number of data services that have been broadly used in big data environments. Archival data must be protected to ensure data integrity, but the cost of data redundancy tends to be much more important than how quickly the data can be recovered in the event of a failure. RAID options have tended to be more widely used in latency-sensitive environments, whereas erasure coding has been more broadly used for data protection in big data and other secondary storage environments. Storage efficiency technologies like compression and deduplication can be extremely important for certain secondary workloads like backup and less important for others because the data is not very reducible. Thus the ability to apply storage technologies selectively is required. Encryption can be important to meet regulatory and compliance requirements for certain data sets, and other features like versioning, audit trails, and WORM may be important in certain archival environments. When geo-distribution of data is required, asynchronous replication is also of interest as a space-efficient way to synchronize remote data sets as things change. Figure 2 illustrates the areas where big data originates.

FIGURE 2

Sources of Big Data



Source: IDC, 2018

Businesses will be collecting and retaining much more data from a variety of sources than they have in the past, driving a very large big data market.

Big data environments tend to support much larger data sets, so infrastructure density is an important consideration. IT organizations are willing to trade rapid recovery times for less expensive but still rock-solid data integrity over long time horizons. As flash costs continue to drop, the industry is likely to see more scale-out all-flash platforms, but HDD-based platforms will thrive in this arena for a long time to come. Configuring the right platform for big data environments is all about meeting less stringent performance requirements while optimizing for easy scalability, manageability, and low cost. Certain workflows against massive data sets targeted for long-term retention that will not be accessed very frequently require higher performance, and small cache layers built from higher-performance media can be a cost-effective way of meeting that requirement. Flash cache tiers can also help complete data movement operations more quickly, staging the data first to the higher-performance flash media to free up the source and then asynchronously migrating that data to the more cost-effective high-capacity tiers on the back end. Flash media, regardless of whether it has SCSI or NVMe interfaces on the device, provides much better data mobility — a feature that will be important for the workflows in many big data environments.

Big data workloads include legacy secondary storage applications such as backup, disaster recovery, and archive, as well as newer big data and analytics applications built around Hadoop, Splunk, and other similar platforms. Traditional business analytics have been more batch oriented, but with the advent of big data this is changing significantly. Real-time big data and analytics is in fact driving a new set of storage infrastructure requirements that run on top of unstructured data sets, resulting in the definition in this paper of big data environments. The ability to configure higher-performance storage tiers on top of what otherwise might look like a massively scalable big data workload will be key to meeting these types of real-time requirements. An example of a real-time big data and analytics application comes from the retail industry. A major shoe manufacturer monitors social media in real time around a worldwide sporting event, such as the Olympics or the World Cup, and uses preexisting customer profiles to generate time-sensitive offers that are tied to the action in the sporting event (which could in the Olympics, for example, be only a 30-second event like a swim heat). The backing data sets are huge but provide key information that allows offers to be individually tailored, resulting in a high response rate from potential customers. This application was custom built by the retailer, depends on high-performance NVMe technologies as well as others in a tiered storage environment, and directly drives revenue generation for the retailer.

FUTURE OUTLOOK

In the legacy hardware-defined era, IT infrastructures typically included many different types of platforms to meet different workload requirements. These systems could be statically configured for a single workload type but had difficulty running mixed workloads with much efficiency. With many different "storage silos," each of which required its own set of administrative skills, this infrastructure was not only complex to manage but also very expensive and inflexible. The importance of software-defined infrastructure, typified by virtual platforms, is that it enabled significantly improved efficiencies in IT resource utilization.

In the vision for the software-defined datacenter, IT resources can be independently allocated as needed to match workload requirements yet support centralized management of the entire configuration through a high-level management toolset that supports heterogeneous storage. Logical

"servers" can be dynamically created by selecting the desired amounts of each resource, configured to support the data services required for that workload, and made available to end users very rapidly. Once that "server" is no longer needed, it can be dissolved and its resources added back into disaggregated resource pools until they are reused with another server. This is the vision of truly composable IT infrastructure, which at this point while still in the future will run on top of fast data and big data storage resources.

As the fast data and big data workload classes evolve, the definitions of the right platforms to service them also evolves. The underlying platforms must support the right data types (block, file, and/or object) and provide the right resources and data services to allow appropriately configured fast data or big data platforms to be dynamically assembled as needed and easily evolved over time. This has implications for customers buying new storage platforms to service these needs in the coming years. Customers should look for systems that support the right resource type and capabilities for either fast data or big data, are software defined for flexibility, and support nondisruptive migration paths to next-generation technologies. Over the next year or two, many datacenters will be built around storage infrastructures that offer both fast data and big data platforms, which can be dynamically assigned to relevant workloads.

CONCLUSION

Next-generation workloads in the mobile computing, social media, and big data and analytics arenas are driving a new set of storage infrastructure requirements for the future. Storage platforms will be much more software defined, and the agility to dynamically configure IT resources into logical "server" definitions will enable much more efficient and responsive IT. The new set of storage requirements is bifurcating into fast data and big data solutions, and by understanding these evolving platform definitions, IT organizations will make better choices upon technology refresh going forward.

The fast data and big data definitions go beyond the traditional primary and secondary storage workload definitions because of the evolving mix of applications that enterprises will be supporting into the future. Key to these platforms is that they are flash optimized for high performance and efficiency and that they exhibit the flexibility to support different media types in tiered storage environments. Scale-out architectures, comprehensive enterprise-class data services that can be selectively applied at the application level, and high availability are requirements across both platform types, with more performance and rapid recovery capabilities necessary for fast data environments and more storage density, massive scalability, and low-cost features required for big data environments. Fast data and big data storage platforms should support excellent datacenter integration capabilities and be easily adaptable to use in the composable storage infrastructure future that will develop over the next several years.

As enterprises undergo digital transformation and look to replace legacy primary and secondary storage solutions, the use of purchase criteria driven by an understanding of the quickly emerging fast data and big data platforms will help inform better business outcomes. Enterprises should start to think now about the implications of these fast data and big data tiers and how they might evolve their own IT infrastructures in this direction.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com