White Paper

Universal Storage for Data Lakes: Dell EMC Isilon

By Nik Rouda, ESG Senior Analyst; and Terri McClure, ESG Senior Analyst

November 2016

This ESG White Paper was commissioned by Dell EMC Isilon and is distributed under license from ESG.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

Contents

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

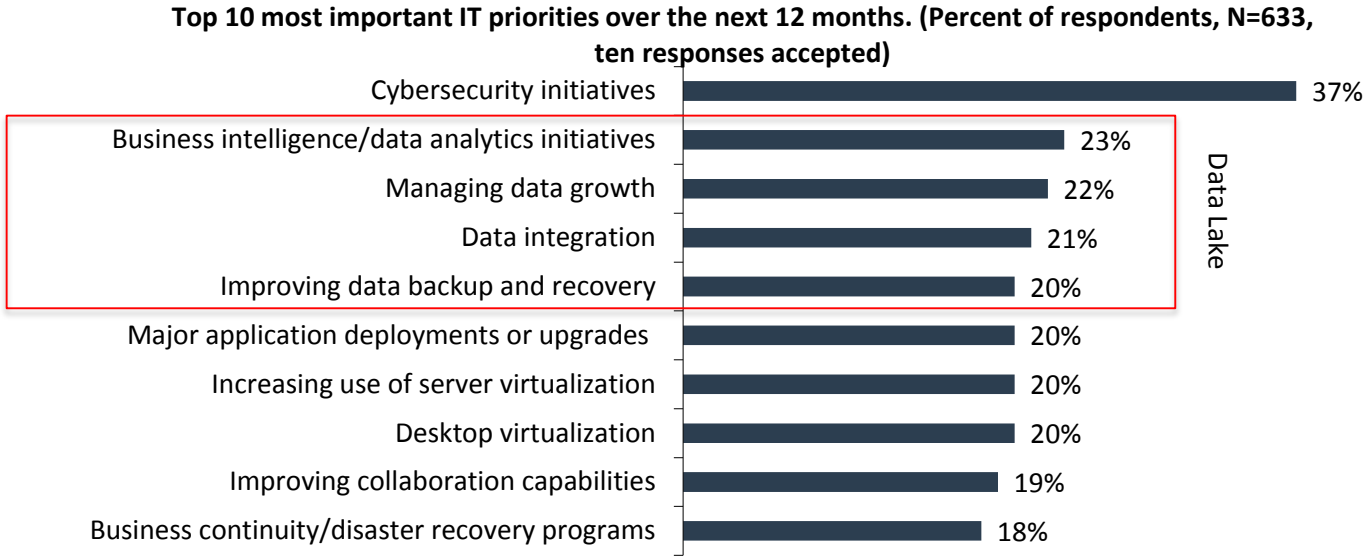
Big Data Needs Big Storage

A rising tide of information is being collected, processed, and analyzed by enterprises around the world, in both traditional data warehouses and newer data lakes. Yet this flood of data brings as many challenges as it does opportunities. As companies become more data-driven in a wide range of activities, they will need their big data storage infrastructure to meet strict enterprise production requirements such as scalability, performance, availability, security, and compliance. At the same time, many organizations are looking to modernize their data centers by implementing enterprise data lakes to increase efficiency and extract more value from their big data assets. Often, consolidating data into a shared storage infrastructure is the first step in implementing a data lake. In addition to supporting multiple workloads, the data lake enables organizations to leverage various big data analytics engines to gain new insight and support business decision-making. For a data lake in particular, the quality of the storage systems can make or break an implementation. Dell EMC Isilon is a leader in scale-out storage and offers many advantages for big data analytics, built on the company's years of experience in large enterprise data centers.

Businesses Want Data Lakes to Have a Big Impact

There has been no shortage of press on the many practical applications of big data in all industries and across all lines of business. Many of these stories are compelling anecdotes, and are often specific to the particular organization's goals and activities. However, some common trends can be found across industries. ESG research has shown that 50% of respondents responsible for their organization's big data and analytics initiatives rank this as their single most important IT initiative for the business.[1] A broader study across all IT decision makers puts business intelligence and analytics (23%) as second only to cybersecurity (37%) as an IT priority, and virtually tied in importance with the related challenges of managing data growth (22%) and data integration (21%).[2] Even data backup and recovery (20%) is relevant, since an important advantage of an enterprise data lake is to provide consistent levels of data protection and security that meet corporate IT requirements. Achieving this in a siloed environment is often a significant challenge. This is especially an issue for remote and branch offices that generate and manage data. These efforts are all in play around the concept of an enterprise business data lake.

Figure 1. Top Ten Most Important IT Priorities for 2016



Source: Enterprise Strategy Group, 2016

[1] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux*, July 2016.
[2] Source: ESG Research Report, *2016 IT Spending Intentions Survey*, February 2016.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

Implicit in all these goals is the need to not just serve the business with *more data*, but to do so with *greater efficiency*, *increased operational flexibility,* and *more robust and timelier analysis*.

Extending the Data Lake Beyond the Data Center

While most would think of a data lake as a centralized cluster of Hadoop nodes co-located in a data center, this definition is shifting somewhat, pushed by the realities of the enterprise. Businesses have significant quantities of data in multiple data centers, in remote branch offices, and indeed in public cloud services, too. This geographic distribution of data is challenging the standard definition, and causing many to look for solutions that don't require the massive overhead of pulling all data back across the wide-area network (WAN) to be redundantly held in the data center for inclusion in analytics. It is in fact possible to have a virtual data lake that spans many distinct environments, with central control and ubiquitous analytics spanning the world.

Similarly, public clouds can act not just as application data sources for data lakes, but also as cost-effective archives for infrequently used data. Private clouds can bring similar benefits, too, with the added bonus of more direct control of the data for security, privacy, and governance. Depending on the circumstances of your IT skills and budgets, some private clouds can also rival the massive public providers for cost-efficiency. An area for exploration is tiering of these various pools of data to meet specific data and analytics demands, balancing workload performance versus sphere of control versus the projected total cost of ownership over the lifetime of the data lake.

Storage Selection Criteria for Data Lakes

As noted, the choice of storage platform underpins the overall efficacy of the technology stack, and will have ramifications that must be carefully evaluated. There are a number of factors to consider in deciding how appropriate a storage platform is for a data lake environment, including:

- Scalability and efficiency impact the solution on multiple fronts: economic and technical. The system's scalability and efficiency characteristics will both reduce the total cost of ownership (TCO) and have an impact on the ability to ingest and store data. Particular attention should be paid to mechanisms that reduce total footprint, such as overall storage utilization, deduplication, compression, storage tiering to optimize storage resources and lower capital costs, and effective integration with cloud storage resources. Human capital required to manage the system, especially in large or rapidly growing data environments, should also be analyzed in the efficiency category because organizations cannot afford to continue to add staff to manage the environment as data grows. TCO matters as the big data initiative benefits are weighed against both capital and operating expense, including maintenance, support, footprint, and human capital. A reduced cost structure should lead to more data stored (because organizations can now afford to) and more valuable insights realized (as a benefit of having more data to analyze).

- Performance seems like an obvious requirement, but it can be elusive as more users do more comprehensive analysis with larger data volumes. Along with the ability of the storage platform to scale performance as needed, it's also important to recognize that not all applications and workloads require the same level of performance. Also, the value of specific data sets often changes over time. With these points in mind, a data lake storage infrastructure that provides varying performance tiers to deliver the appropriate level of performance needed by specific workloads can help to minimize overall storage costs while meeting the needs of the business.

- Data protection, security, and governance are mandatory for big data environments. As data lakes or data hubs start to encapsulate all manner of data in one central location, this clearly needs to be treated with great care. Data availability and overall solution resiliency become much more important as more people rely on access to

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

this central pool of data. Compliance with relevant government and industry regulations must be addressed directly and explicitly.

- **Accessibility by multiple application types** may be one of the least recognized attributes of the storage decision, but it can provide significant advantages in flexibility of models for enabling different groups or tools to harness the data without moving it into other platforms before processing can begin. Because of this, access controls also must be well developed and granular.

- **Data analytics** must support the use of unstructured data analytics platforms like Hadoop and Splunk. The advantage of a data lake is minimizing or eliminating the need to replicate and move large data sets, which in turn helps to accelerate time to insight and lower network costs. This also eliminates the need for a separate, dedicated analytics storage infrastructure, which helps save on capital expenses, and operational and management costs. Another point to consider is that consolidating data into a centralized data lake simplifies data analytics initiatives, as it eliminates the challenge of obtaining data from various silos that may extend across the enterprise. Not least, the data available for analysis is more extensive, which may yield more valuable insight and results.

Data Lake Storage Alternatives

Again, a range of traditional options for storage platforms include: commodity direct-attached (DAS), storage area network (SAN), and network-attached storage (NAS). Conventional wisdom has been to use commodity storage in the form of internal drives. This DAS approach relates to traditional Hadoop deployments that are typically separate, dedicated "silos" to which data must be copied and moved for analytics projects. As mentioned earlier, this is inefficient and can slow the time to insight. Instead, a full 23% of decision makers want their big data deployment strategy to be on-premises ***with shared infrastructure***.[3] But when weighing the impact of storage infrastructure choices on data management, time to data accessibility (ETL) and analytics conventional wisdom fall short on delivery. The advantages of a "scale-out" storage architecture approach is that it can expand easily in large and growing data environments. This is in stark comparison to traditional "scale-up" approaches that become increasingly complex to manage as data environments grow.

Data Analytics and Data Lakes

For many companies, the time needed to get an answer is the key criterion for the adoption of data-driven decision-making. No longer do quarterly batch reports meet the needs—instead, daily updates, real-time alerts, and ad hoc querying are becoming standard requirements for both business analysts and executive leaders.

Accordingly, vendors are now bringing a breadth of unstructured data analytics platforms into play, including Splunk and various Apache Hadoop distributions that can be applied to a centralized data lake. These may also be supplemented by data discovery, advanced analytics applications, data visualization, and business intelligence tools. With all these software tools come accompanying options in deployment models, including commodity servers, ready-made appliances, or cloud services, and open source or proprietary software. Each choice will impact the overall capabilities of the solution, affecting end-user perceptions of performance, flexibility, and availability.

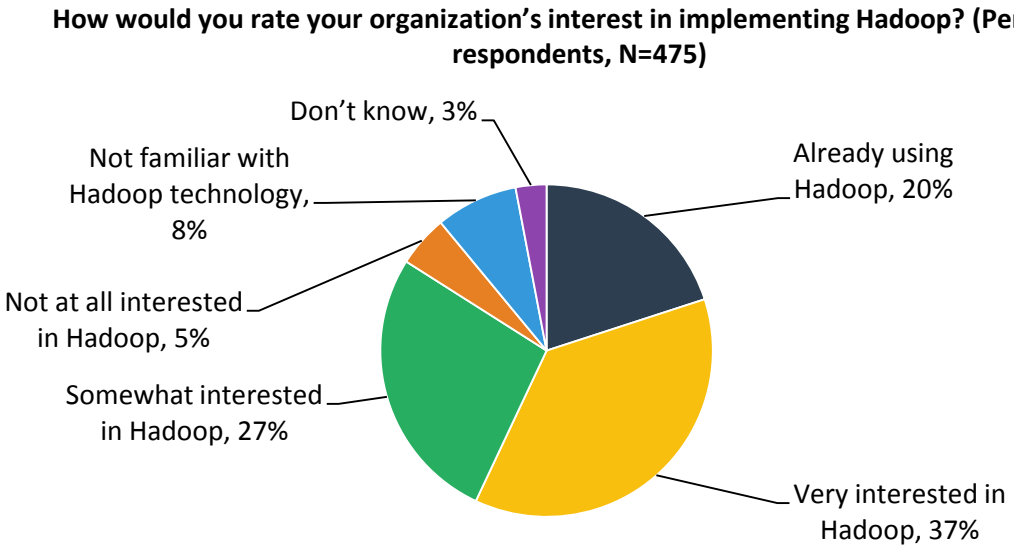
The high expectations of the business put a lot of pressure on enterprise IT departments to deliver a well-implemented solution. This isn't usually an easy task, considering that data lake initiatives often involve the integration of many new data sources, data platforms, and analytics applications with existing data warehouses and transactional databases. One of the core components for a data lake is Hadoop itself, and Figure 2 shows the current rate of adoption, with 20% of respondents in production now, 37% very interested, and 27% somewhat interested. Significantly, more than a third (35%)

[3] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux*, July 2016.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

say they are implementing Hadoop expressly to establish a centralized data lake or hub. A quarter see Hadoop as potentially *replacing* their existing data warehouses in the future while another third intend to *optimize* their data warehouses, with workloads placed in each according to specific needs.[4] This all means that there are already many important data lake deployments in operation today, but many more still to come.

Figure 2. Hadoop Adoption



Source: Enterprise Strategy Group, 2016

The architectural complexity of a data lake spans many IT disciplines, with dependencies on everything, including applications, servers, networks, and storage. The potential storage issues are sometimes glibly underestimated, with the assumption that the Hadoop Distributed File System (HDFS) provides cheap and cheerful provisions for storing and managing massive volumes of big data. The truth is that storage requirements for the enterprise are becoming increasingly demanding, especially as more decision makers become reliant on big data insights. Significantly, 77% of respondents say it is important or crucial for the storage team to be involved for big data success.[5]

Advantages of Isilon Scale-out Storage for Data Lakes

Today, there is still a relative immaturity of functionality and robustness in many data lake technology stacks when it comes to storage. Although Hadoop and HDFS can simplify the model for scaling on commodity servers with DAS, some alternatives provide compelling advantages for the enterprise and help overcome some of the challenges associated with using the traditional approach.

Challenges with using the embedded storage/DAS approach include data protection, data leverage, elongated business process, and, surprisingly, cost. On the data protection front, HDFS uses multiple copies of data to provide data protection, meaning it consumes a lot of storage. Both data leverage and business processes are impacted by the fact that data is only accessible via HDFS and is not accessible to other applications that require other interfaces (i.e., RESTful object-based applications or NFS/CIFS/SMB file-based applications). This means ETL operations need to be performed to ingest or leverage data in other business processes, thus elongating those processes each time the ETL process needs to be performed. This also means that organizations must have multiple data repositories for the same data in multiple data formats to support different business processes. So, on the surface, using commodity DAS configurations may sound

[4] Source: Ibid.
[5] Source: Ibid.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

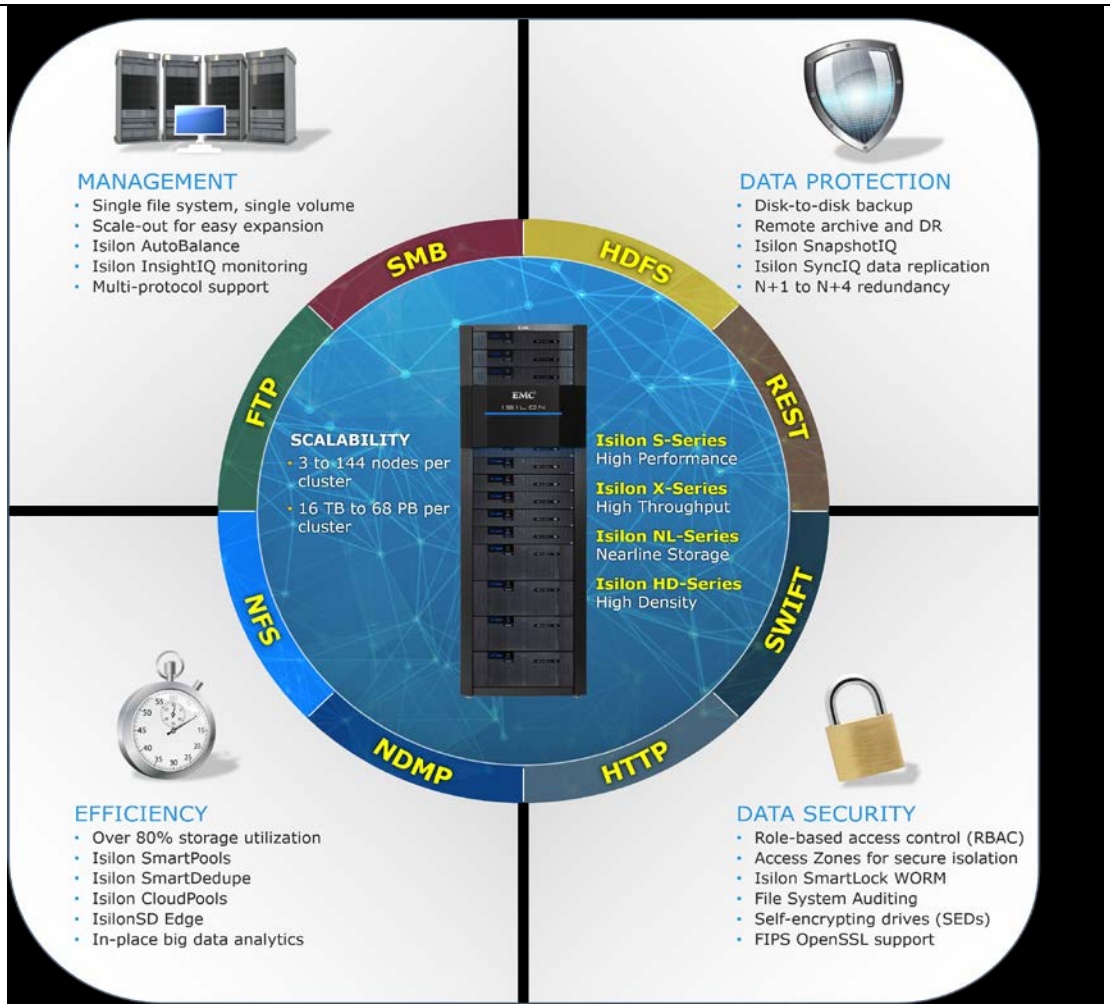
attractive and may indeed be a good fit for many organizations, but those companies that need to analyze data from multiple sources or leverage it to support multiple business processes incur further costs for additional infrastructure and may need to investigate alternative approaches.

One alternative approach that helps overcome these challenges is the adoption of a shared storage platform that has been designed to meet enterprise IT operations requirements. Isilon scale-out NAS storage is a prime example of this case, which brings Hadoop to your data, instead of moving all your data to Hadoop clusters. It lets users create a central data hub that supports multiple applications and business processes, reducing costs and business cycles by eliminating most ETL requirements.

Many Protocols, but Only One Copy of Data

Isilon is a flexible storage platform that supports multiprotocol access to a single data object, eliminating the up-front protocol decision because NFS, RESTful objects, HTTP, FTP, SMB, and HDFS are all supported. So users can ingest an object from a web app and access it via NFS to edit it. Or better yet, a user might access web logs directly from a web application, rather than exporting to a spreadsheet, and access these via the native HDFS interface to run analytics. This ability to make only one copy available for multiple uses is a major benefit for reduction in overall storage costs and cycle time because it means there is no need to export data to multiple systems for the various use cases. A single repository also greatly simplifies compliance audit requirements, rather than requiring user to chase after many distinct locations and sources.

Figure 3. Isilon Data Lake Features



Source: Dell EMC/Isilon, 2016

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

In-place Analytics with Your Favorite Flavor of Hadoop

With native HDFS support, Isilon enables organizations to do "in-place" analytics on data without needing a lengthy data ingest from other primary storage data sources to the Hadoop system, which very often leads to a faster overall time to results. Although more specialized data layouts and approaches can sometimes be faster in querying and analysis, with Isilon, data analysis can be started immediately, and the reduced effort and start time delay without ETL can often overcome the difference. In addition, concurrent instances of different Hadoop distributions could be run in parallel on the same underlying storage system, giving much more flexibility to leverage the relative strengths of each, again, without the need to move large quantities of data around.

Enterprise-class Storage Increases Efficiency and Safety

Although HDFS may be a reliable and scalable model for collecting and storing the high volumes and varieties of data in a typical big data environment, it isn't necessarily the most efficient. Some features that provide that robustness on commodity hardware may actually detract from overall efficiency. Mirroring with Hadoop direct-attached storage is a good example, causing typically three to five times redundancy, which significantly affects the effective usage ratio of total drive capacity. Isilon, with built-in data protection, high availability, and general robustness, can instead run at 80% utilization levels of capacity (compared with 20-33% with HDFS) and this is further improved by data reduction of up to 30% with Isilon SmartDedupe data deduplication software.

Note, too, Isilon's ability to scale from 16 TB to 68 PB in a single cluster, while the OneFS operating system's single file system, single volume architecture greatly simplifies data and storage management. Separating server and storage by growing each independently instead of always adding another fixed unit commodity server also allows more targeted scaling of the environment to meet the actual workloads. All this helps reduce the storage footprint, bringing associated cost reductions in energy and space consumption in the data center.

From a governance and security point of view, the Isilon storage system offers "write once, read many" (WORM) compliance for archival to meet government and industry regulations, standard Kerberos authentication, and access control lists (ACLs) to make sure the user touching the central data hub is authorized.

Recognizing that the next-generation data lake is likely to be a mix of a centralized data lake and ponds—both tributaries and reservoirs—that play specific roles, Isilon supports edge and cloud deployments with its software-defined storage architecture to meet these needs.

Including Remote Data with Isilon on the Edge

IsilonSD Edge is the tributary that extends the data lake to enterprise edge locations, including remote and branch offices, allowing IT organizations to centralize data analytics operations, simplify management, and protect unstructured data used to support a wide range of workloads. With SyncIQ software (included with IsilonSD Edge), data can be replicated from remote and branch offices back to a central location for analysis in near real time. And this replication eliminates the need for edge-based data backup for unstructured data, helping to further drive down costs.

SyncIQ can also be used to distribute data in near real time, so upon completion of analytics jobs, results can be shared quickly, keeping all offices up to date with the most recent information.

Leveraging Isilon CloudPools

Isilon CloudPools allows enterprises to archive or tier data from an on-premises Isilon cluster to a choice of cloud service providers or even a private cloud. It acts as a bottomless pool of capacity for archive or tiering of stale or cold data, yet

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

keeps that data visible and accessible for reuse. It allows the secure leverage of cloud-based resources with built-in data encryption; the cloud service provider cannot read data stored in an Isilon CloudPool.

Isilon's ability to provide policy-based, automated storage tiering (with Isilon SmartPools) within the data center and with Isilon CloudPools to tier data to a choice of public or private cloud services is especially useful for storage of "cold" or "frozen" data. All of these features combine to reduce initial cost of purchase, ongoing operational costs, and risk of failure or security breach of sensitive information.

The Bigger Truth

Having explored the rapid growth of big data in adoption and importance, and the potential impacts of the underlying infrastructure, it is clear that enterprises should rethink the architectural implications of their storage choices for their data lake initiatives. There are multiple advantages in taking a shared storage approach, covering a wide range of desired characteristics, including increased efficiency, reduced total cost, overall speed to answer, reduced risk of data loss or inappropriate access, and analytics flexibility.

Isilon is breaking ground in challenging the default storage paradigm assumptions of big data vendors, and its approach is well worth evaluating for its merits compared with the de facto standard of direct-attached storage in commodity server hardware. Coming from a long history of building flexible, scalable storage platforms for demanding enterprise requirements serves Isilon well in addressing many common challenges of data lake storage, and this experience should serve customers well. Particularly, current Isilon customers should experiment with running Hadoop on their existing systems; they may well find that the right answer is already in place.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of The Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at 508.482.0188.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.

 www.esg-global.com

 contact@esg-global.com

 P. 508.482.0188