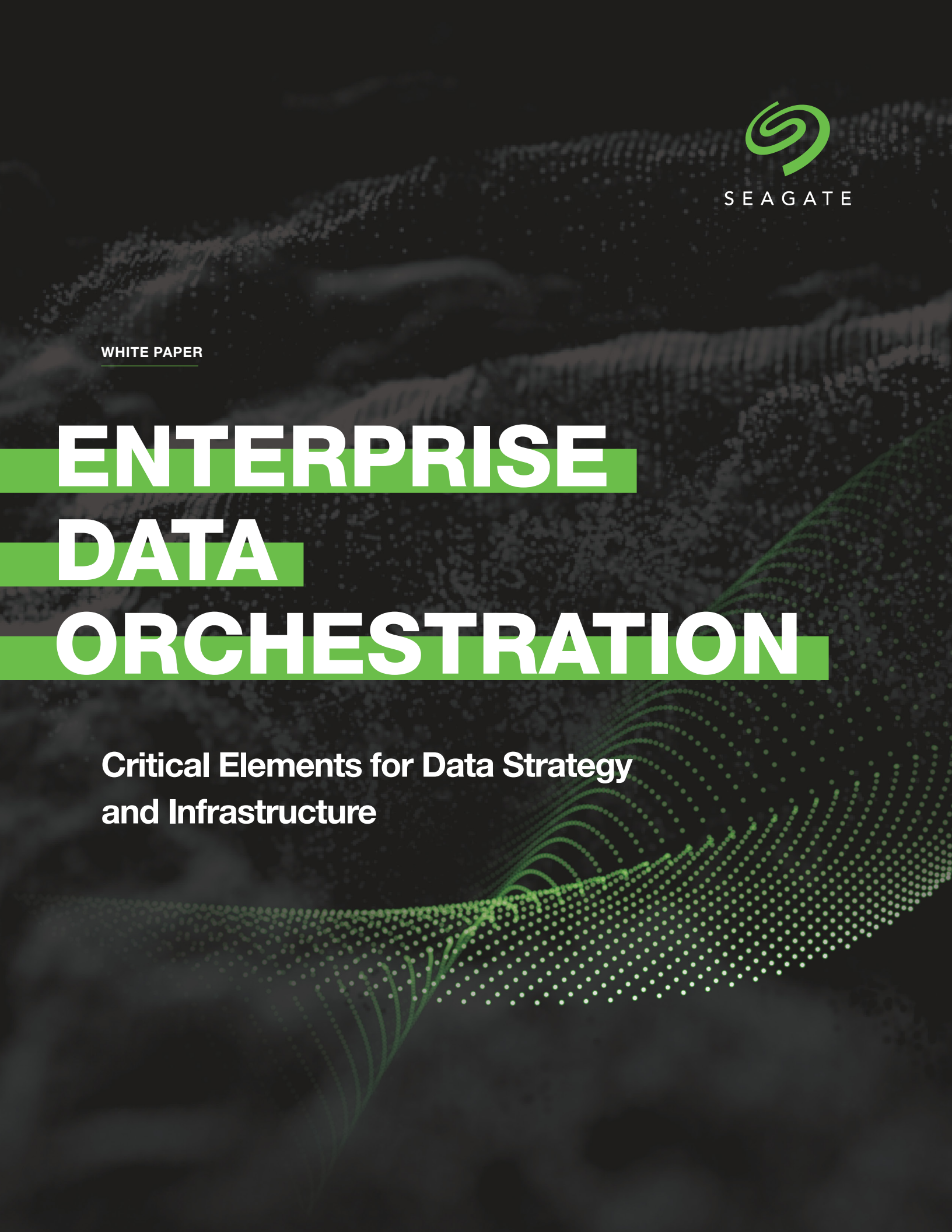

SEAGATE
WHITE PAPER
ENTERPRISE DATA ORCHESTRATION
Critical Elements for Data Strategy and Infrastructure

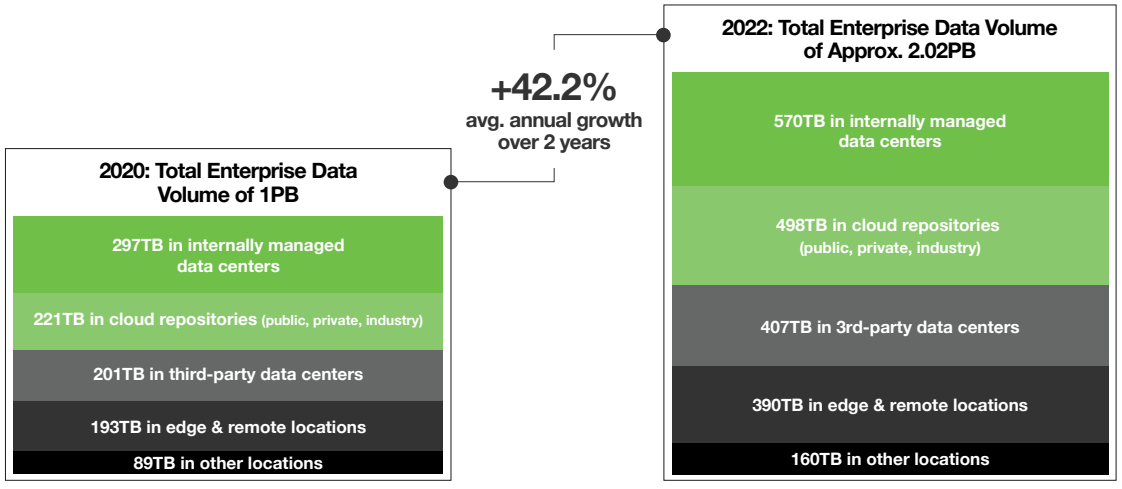
CONTENTS

Introduction

Managing enterprise data has never been more complex, and trends indicate this complexity will continue to grow. Data is now an essential asset, similar to physical capital and intellectual property. As the quantity and importance of data increases, so too does the complexity of managing it—especially when the data is distributed all over the place, from endpoint to edge to core and cloud data centers. With data distributed everywhere, the modern distributed enterprise requires new methods for data movement and orchestration.

According to Rethink Data, a Seagate report with research and analysis by the International Data Corporation (IDC), enterprise data is now proliferating from myriad sources at an unprecedented 42.2% annual growth rate. This growing data is also sprawling through various IT configurations, which includes multiple levels of public, private, and hybrid cloud—a multicloud architecture. The increasingly intricate movement of data in an increasingly varied ecosystem compounds data management challenges for business owners.

FIGURE 1 **Expected Annual Data Growth Rate**



Source: The Seagate Rethink Data Survey[1], IDC, 2020

To complicate matters further, the frequency at which enterprises need to move data is also increasing. IDC says the volume of data that enterprises periodically transfer from edge to core will grow from 36% to 57% within the next two years. Meanwhile, data that's transferred immediately from edge to core as soon as it's captured will double from 8% to 16%. This means enterprises will have to manage more data in motion than ever before.

Equally crucial is the fact that data of all kinds now has much greater potential value than it has had in the past. That's because advanced analytics, artificial intelligence (AI), and machine learning (ML) can discover, glean, and develop the value of data in transformative ways. As such, it's critical that each enterprise maintain, track, and use all its data. But according to IDC's research in the Rethink Data report, enterprises today only put 32% of available data to work—the remaining 68% goes untapped.

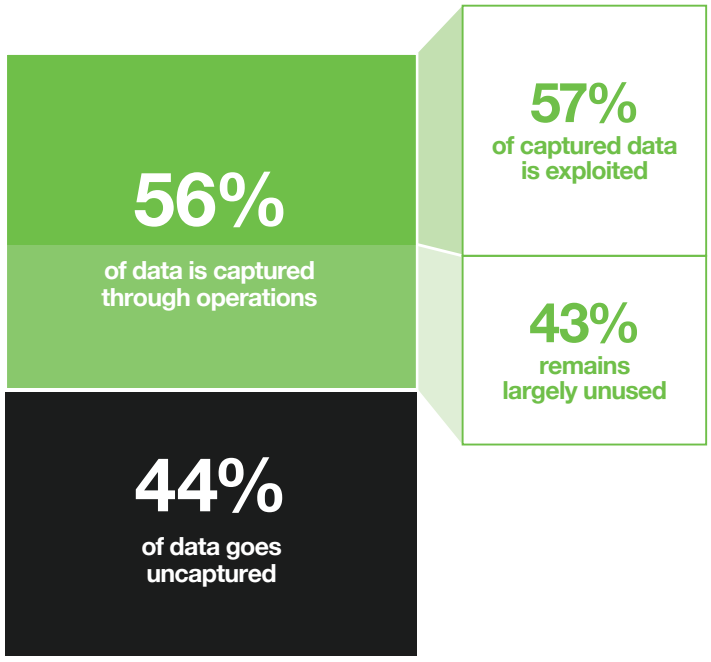
Data Orchestration Must Become More Sophisticated

Most enterprise IT decision makers know they need to implement DataOps—a discipline that's focused on connecting data creators with data consumers. Today only about 10% of organizations report having implemented DataOps fully across the enterprise. But according to IDC, deploying DataOps along with other data management solutions will lead to measurably better business outcomes, including boosted customer loyalty, revenue, profit, and more.

Executives and other decision makers should consider all dimensions of the data life cycle and how they impact an organization's ability to effectively utilize its data. Organizations especially need to understand how data is used, how data moves, and how to orchestrate data at scale so it's available, protected, available for analysis, and ready to be put to work.

Failure to adequately address these factors can adversely affect an organization's ability to leverage one of its most valuable resources.

FIGURE 2 How Much Data Actually Gets Put to Work?



Source: The Seagate Rethink Data Survey, IDC, 2020

How Data Is Used: The Data Life Cycle

Data has a life cycle in the sense that what one does with data changes over time. Data is also in motion, moving from its point of origin to initial ingestion platforms and through to storage and analytics systems. Data requires protection across its full life cycle and across the various kinds of compute, storage, and network infrastructure that operates on that data. This occurs over four phases.

Origin Phase

The origin phase of the life cycle is the point at which data is created. There are many diverse sources of data with distinct characteristics, and each of these must be accommodated. For example, the sensors on manufacturing equipment or autonomous vehicles can generate enormous amounts of data. Transmitting all that data to a centralized processing system may not be practical. Rather, data captured at these endpoints might be better processed locally or at the edge, and the results of that initial processing could be sent to core, centralized systems either via network wires or via physical storage devices if data quantity is enormous and speed is critical.

Alternatively, there may be use cases in which all data should be captured. In these situations, data could be transmitted to scalable, low-latency data stores. Applications that require all data be captured but do not require immediate access to the data may be better served by using batch uploads to a private or public cloud.

Ingestion Phase

The ingestion phase occurs when data is acquired and brought under the control of an organization. Salient issues during this phase include the volume of data, the rate of ingestion, the security of data, and the processes for validation. In some cases, data may be cached locally and transmitted in batch. For example, a stream of time series data may be analyzed locally to detect anomalies in the data stream as soon as possible. The data could then be transmitted, either in streams or in batches, to a persistent storage system. In other cases, ML models—which are used to perform preliminary validation and analysis and transmit derived results—may be deployed to the edge.

Analysis and Consolidation Phase

During the analysis and consolidation phase, data is used as part of a business operation. This can range from executing transactions to training ML algorithms. In the past, large data sets were of limited use, but the application of ML in business has changed that. The quality of ML models depends on a combination of algorithms and data. In many cases, models can be improved by training them using more data. Just as humans can learn about exceptions and unusual instances by experiencing more, ML models can discern infrequent patterns that influence the model's predictions. One of the most difficult—if not impossible—challenges in data management is to determine what data will be useful in the future. Data that is not apparently valuable now may be valuable down the line. For example, self-driving cars were having trouble identifying pedestrians wearing yellow shirts

when making turns. ML engineers used metadata-tagged images from existing data sets to find additional examples of such instances and used them to further train the model.

Since it is essentially impossible to know what data will be useful in the future, and data is used to generate significant intellectual property in the form of ML models, organizations are well justified in saving virtually all their data.

Archiving Phase

The fourth and final phase of the data life cycle is archiving. This stage must be managed through a comprehensive orchestration platform—one that spans infrastructure from data centers to clouds. In the past, a collection of best-of-breed devices and storage software may have been the best solution; but that's no longer the case. The components in these ensemble solutions each tend to locally optimize a part of the life cycle, but that does lead to an optimal global solution.

Data Movement Across the Enterprise

Organizations should examine how data moves across devices and deploy a solution that accounts for orchestration across all infrastructure. There are several classes of storage infrastructure to consider.

Endpoints are the origination of much of the data an organization is tasked with managing, which includes preliminary analysis. Endpoints include a diverse range of components, including mobile devices, internet of things (IoT) sensors, and autonomous devices.

That said, not all data generated by edge devices is directly transmitted to a data center or cloud. Some filtering, aggregation, and preliminary analysis can occur closer to the point of data generation. This practice is known as edge computing.

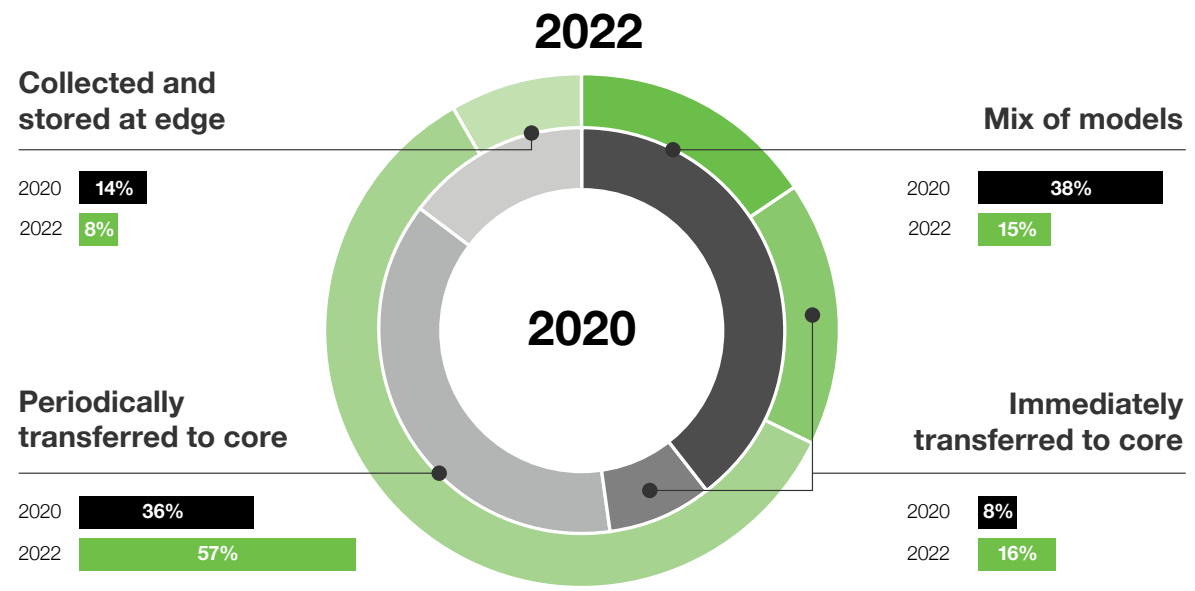
While edge computing offers an advantage in allowing for immediate data analysis, it does introduce additional management responsibilities. Local processing will be useful when applications demand real-time decision making. But for deeper analysis, AI, and ML practices, training the models must still take place at the core at scale. So, to ensure data is utilized as needed and understood well enough to be utilized more effectively, data managers need to move much of the data to the core to continue improving analytical models. Thus, storage administrators must consider the simplest and most efficient way to secure and move data from the edge.

Relatively small amounts of data can be efficiently transmitted using 5G networks, but the cost is prohibitive for larger volumes of data. Wired networks offer lower cost and higher capacity but may not be available to all endpoints or edge devices.

Data centers are hubs of compute and storage services, and the place where much of an organization's storage infrastructure is located. These data centers are increasing complemented by public and private cloud storage. The combination of on-premises and cloud resources are known as hybrid clouds and are becoming increasingly popular—particularly for storage.

Clearly, data moves in a variety of ways across enterprise infrastructure from the edge, through ingestion pipelines and analysis workflows that may be in data centers or public clouds. Managing this flow requires orchestration software that spans infrastructure from the edge to the cloud.

FIGURE 3 **Approach to Collecting Data at the Edge Now vs. Two Years From Now**



Source: The Seagate Rethink Data Survey, IDC, 2020

Requirements for Data Orchestration at Scale

One of the first requirements of data orchestration is a recognition of the scale at which it must function.

Enterprises need ways to efficiently move petabyte scales of data across storage systems. Attempts to move data at this scale using networks would consume virtually all throughput for extended periods. Moving just 20TB of data with a 100Mbps upload speed would take more than 20 days. Even with gigabit-speed optical-fiber service, assuming an average sustained upload speed of 800Mbps, it would take 2.5 days to upload 20TB—and almost a week to upload 50TB. Networks are not fast enough or cost-effective enough to regularly move terabyte- or petabyte-scale data sets.

Additional infrastructure to enable data in motion is required to complement networks. This complementary infrastructure should include storage devices, such as physical data shuttles, that can quickly load data before being physically shipped to a data center. The use of variable paths and processes to move data, in turn, requires an orchestration layer that's capable of choreographing a complex flow of data. Getting data from where it is to where it needs to be requires policy-based management. It would be virtually impossible to try to manage the flow of data by directing individual data sets or workloads. Instead, policies must be defined that specify how data should be moved and stored based on the attributes of that data.

Defining Policy-Based Data Management

Policies specify several things about how data is moved and stored.

For example, data in motion should be encrypted in many cases. This requires the use of a particular encryption algorithm and an encryption key. Encryption keys, in turn, have their own life cycle that is typically managed by a key management service. Data at rest often requires encryption as well, and that drives the need for additional specifications within data management policies.

Different sets of data have different latency requirements. Some data has low latency requirements and must be analyzed shortly after it is generated. IoT sensor data that is monitored for anomalies, for example, should be analyzed as soon as possible in order to detect potential problems. Credit card transactions are another example in which immediate analysis is required to mitigate the risk of fraud. But data management doesn't end there for these data types. Following such immediate analysis, data from these sources retains significant value for developing deeper understandings—analysis of transaction interrelationships and long-term trends, for example. As such, policy-based orchestration must account for follow-on management phases for every piece of data.

Of course, many business applications are not as time sensitive as sensor readings or credit card transactions. Some data can arrive up to 24 hours after generation and still provide full value. A retailer may bulk upload inventory data in the middle of the night to determine the logistics of restocking over the next several days.

Another factor to consider in regard to data orchestration is the volume of data. The volume of sensor data, for example, is a function of the number of sensors and the frequency with which each sensor transmits measurements. The rate of growth in data is determined by the rate at which sensors are deployed. In the inventory example, however, the volume of data is determined by the number of products in the inventory, and the rate of growth is determined by the rate at which new products are added to the inventory.

To summarize, when developing data orchestration policies, map requirements for latency, expected volumes of data, and security controls to methods for moving data.

Additionally, policies will depend on data-set metadata. Metadata is often implemented using tags that enable integration, selection, and filtering through various stages of the data life cycle.

Initial Storage and Processing

The very large existing and anticipated growth in volumes of data in many organizations provides a clear incentive to drive down the cost of storing and moving data. These costs must also be balanced against the opportunity cost of losing data. As more enterprises employ data science techniques and ML models, they discover novel ways to benefit from data that may not have had a previously apparent use.

In highly distributed environments, some data will move over third-party networks, including 5G networks. In those cases, both cost and security should be considered as issues. Third-party networks may be the best choice for meeting low latency requirements, but the cost is higher than other transfer methods. The potential business value of immediate access to data should outweigh the cost of employing low latency but higher cost transmission methods.

Of course, once data arrives at a processing point, that does not mean that is the end of its movement. Data may arrive in a data center before it is moved to a cloud. Data within one cloud may be moved to another cloud, or back to the data center. One of the challenges enterprises face in this situation is the lack of standardized tools for data movement in the public-cloud landscape. There is no cloud-neutral way of moving terabytes of data. Each cloud vendor has its proprietary methods for moving data to and from the cloud. Ideally, enterprises should deploy a data orchestration system that functions within a multicloud environment and can move data in multiple directions across those platforms.

Analyzing Data at Scale

It is well understood among ML practitioners that the quality of a model is a combination of algorithms and data. Since algorithms are well known and well understood, it is data that can often be the distinguishing factor when competing using ML systems. Another well-known fact of ML in practice is that one does not always know what data will be useful ahead of time.

Sometimes testing reveals that an ML model does not perform well under certain conditions. This can be corrected by providing additional training for the model specifically focused on poorly performing areas. In this case, having access to data that is well curated with descriptive tags would enable ML and data engineers to find relevant data for additional training.

Which data, and how much data, are as important to effective analytics as developing effective ML models. Data cannot be effectively deployed in machine learning models if large portions of interrelated datasets are lost. Thus, ensuring potentially valuable data is not lost is a central part of a long-term strategy for developing intellectual property from data.

Data loss can come in a variety of forms. Data can simply be deleted after a period of time. This made sense when storage was scarce and costly, but that is no longer the case. Concerns about the cost of transmitting data over 5G networks might motivate enterprises to aggregate data at edge devices before sending it to the data center or a cloud. Aggregation, however, can add latency to the ability to analyze detailed data that often provides significant real-time value. Consider time series data about a manufacturing machine's performance. Anomalies that are apparent in data that's reported every 5 seconds may be missed when data is aggregated at the edge to the minute level.

The value of data will likely increase over time as organizations discover new ways to use data. In many ways, data is a currency of business value and should be preserved—especially when the cost of preserving data is dropping.

Software Drives Orchestration

The software layer of storage infrastructure is crucial for efficient and effective data orchestration. Software enables important aspects of data management, such as ensuring data protection, minimizing operational overhead, optimizing for costs, and enabling policy-driven management.

A best practice is to leverage open source software and commodity storage systems along with software-defined storage to optimize for costs. This approach mitigates vendor lock-in.

Another best practice is to store data rather than discard it. It's better to have large data sets that can be used for analytics and ML than to reduce storage costs and limit the organization's ability to build quality, effective models.

Data orchestration software optimizes for costs by managing based on policies, which, in turn, is enabled by tracking the provenance of data and other metadata. Also, with well-curated data, organizations can use AI to make decisions about how to ingest, move, store, and process data. Automated AI services that are aware of data's value and function can apply policies as data moves through storage infrastructure. Policies can integrate with data pipeline processing. For example, data managers can set policies to move small or high-value data sets across networks while moving lower-value data via physically shipped shuttle devices. Storage managers can also leverage storage

management applications that do not require virtual private networks or a physical connection to manage a network.

Security Considerations When Configuring Data Orchestration

When establishing a data orchestration system, consider a broad range of data privacy and security requirements.

Compliance requirements for data use, permissions, sharing, privacy, and security vary by industry and data characteristics. This creates additional challenges for meeting requirements in a cost-effective way. Data orchestration can help.

Enterprises can track data integrity with fingerprints or message digests of data. This can help ensure that data is not tampered with. If a fingerprint does not match what is expected, a file can be quarantined for review by a data administrator who can then accept or reject the data.

Also consider who has access to data. If certain data provides a competitive advantage, consider whether certain cloud service providers may have competitive conflicts.

Additional building blocks of compliance include blockchain tracing and provenance tracking. Securing data is a responsibility that spans the full life cycle of data. This requires establishing a root of trust within the data orchestration system. Data provenance information must be consistently collected, stored, and managed. Mechanisms should be in place to ensure data integrity and trustworthiness while also protecting the privacy of sensitive or confidential information.

Conclusion

More data—data that's more distributed than ever before—is a source of competitive advantage; it's imperative that organizations harness this opportunity.

The goal of data management is to facilitate a holistic view of data and enable users to access and derive optimal value from it—both in motion and at rest.

The most important step to deploying a successful data orchestration strategy is to investigate and consider how to develop the IT enterprise infrastructure. In addition to scalability, this infrastructure must enable data to be efficiently and rapidly captured, moved, analyzed, and put to work using distributed storage architecture and a rich software-based control layer that brings data to bear where it's needed most.

Ultimately, an active and efficient DataOps program will rely on effective data orchestration as its foundation for building and training AI models and for deploying analytics at scale. In the end, these advanced analytical results can lead to competitive advantage.

Ready to Learn More? Visit us at **seagate.com**

SEAGATE

Seagate Private Cloud Solutions
www.seagate.com/solutions/cloud/private-cloud/

Lyve Rack – Open affordable, and durable object storage solution
www.seagate.com/products/storage/object-storage-solutions/lyve-drive-rack/

CORTX – Open source mass capacity object storage
www.seagate.com/products/storage/object-storage-software/