SEAGATE | CORTX™

Solution Brief
CORTX™

CORTX: Open Source Mass-Capacity Object Storage

Democratizing Access to Data via Mass-Capacity Data Platform Solutions

Introduction

We are living in a tremendously exciting era of innovation and opportunity, and at the same time, one of unprecedented challenges. Advances in artificial intelligence (AI) are unlocking many opportunities by delivering solutions to critical challenges in almost all fields of human endeavor. From improving our health with custom DNA diagnoses to making the roads safer with autonomous vehicles, AI is poised to transform our world in unimaginable ways. A key driver of AI enhancement has been innovations in machine learning (ML)—the technique by which we train machines to act autonomously—and corresponding innovations in big data analytics, which allow us to find hidden insights in ever-larger data sets.

Our future depends on the ability of ML and big data analytics to extract actionable insights from data. Year over year, the exponential growth[1] in data is what offers the most promise for continued innovation. Research in AI has repeatedly shown that the accuracy of autonomous agents is directly related to the size of the data set on which they are trained, and the quality of data analytics also benefits from larger data sets. A recent study by IDC (Data Age 2025) is encouraging on this front, as it predicts that the amount of data the world will produce will grow from 33 zettabytes (ZB) in 2020 to 175ZB in 2025. With so much data generation, we can eagerly anticipate the tremendous benefits for our world: inches of sea level rise prevented, acres of forest saved, accidents avoided, and additional vaccines created, just to name a few.

Unfortunately, as described in the IDC report, challenges remain to realize all this potential.

1 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Only a fraction of that 175ZB of data is actually going to be stored; the vast majority of that data and the potential value it holds could go untapped. Why is it that data, which is the lifeblood of the digital economy, is so readily discarded?

One critical reason for this is the relatively high cost of storing data. If the cost of storing data could be significantly reduced, then more of that 175ZB would be stored, which in turn would lead to improved insights into our world and the critical discoveries we need.

This analysis is visualized in the below graph (figure 1), which shows the tremendous amount of growth in data creation and the relatively small amount of this data that is actually being stored.

FIGURE 1

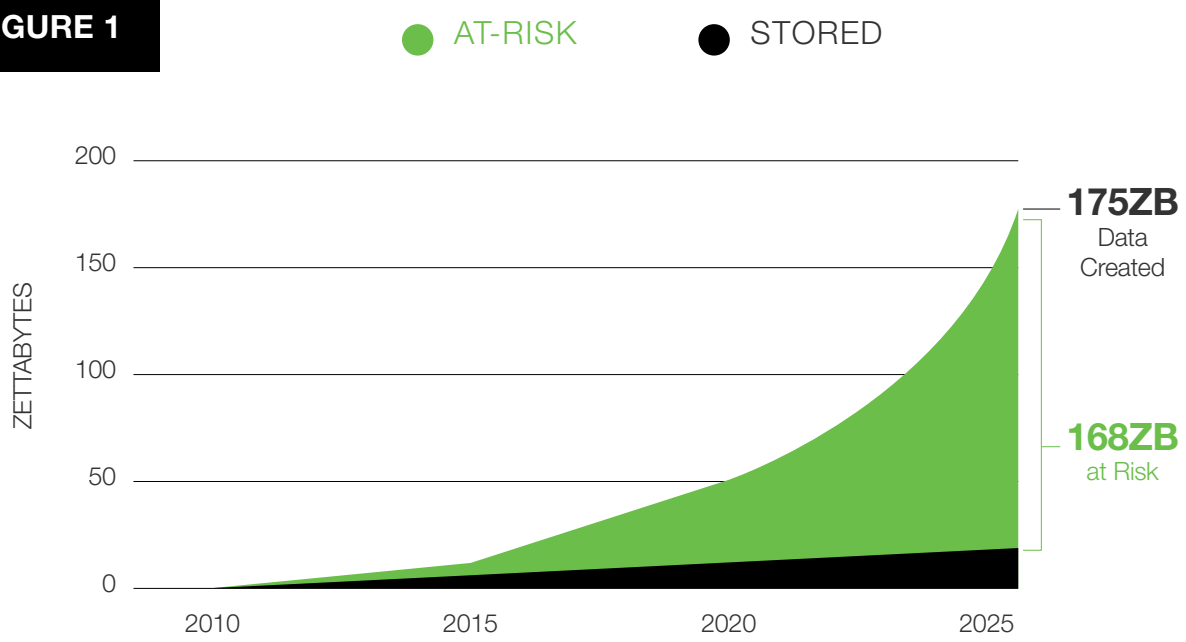


Figure 1: Stored (black) vs lost data (bright green) data

In this graph, the top right edge of the slope shows the total amount of data that IDC predicts will be created. The black sliver at the bottom is the relatively small amount of data that IDC predicts will actually be stored. Per the graph, the amount of at-risk data—i.e., data that IDC predicts will not be stored—dominates the relatively small amount that's predicted to be stored. As discussed, each additional byte of data creates more accurate autonomous AI agents and higher quality insights from big data analytics.

This untapped data represents a tremendous amount of unrealized potential

From market analysis done at Seagate and at IDC, this massive amount of data and potential will go unrealized due to the relatively high cost of current on-premise and cloud-based storage solutions.

To address this, Seagate is improving economic efficiencies of both hardware and software. Hardware innovations, such as heat assisted magnetic recording (HAMR) and multi-actuator hard drive (MACH)[2], significantly lower the cost of storing and accessing each byte.[3]

However, these new hardware innovations alone are insufficient to address the problem. To fully realize their benefits in the data center requires a completely new data platform designed with and for these devices. Seagate anticipated this problem, and we are addressing it with the introduction of **CORTX**: a mass-capacity object storage data platform co-developed and co-designed with Seagate's best-in-class mass-capacity storage devices.

Seagate is making CORTX open source with a view to making hyperscale storage capabilities accessible and affordable for all businesses and organizations. Doing so will ensure that the improved economics of our mass-capacity devices are delivered efficiently to end users while saving at-risk data.

CORTX Open Source Object Storage

CORTX is an open-source software-defined object store backed by Seagate and designed, built, and maintained by a growing community of data scientists and big data and enterprise storage experts.

Designed by a consortium of high-performance computing (HPC) experts with an eye toward future data center requirements, CORTX enables maximum scalability, resiliency, and hardware efficiency. Without any concurrency-reducing global locks, CORTX provides immediate consistency for object access. For maximum scalability, concurrency, and searchability, CORTX distributes metadata management across all servers. To ensure data resilience for ever-larger data sets, CORTX erasure coding can provide multi-tiered data protection across all known data center failure scenarios.

Uniquely, CORTX benefits from a tighter integration with storage. CORTX enables the data to communicate directly with the storage drives without an intervening file system. This direct-to-drive architectural feature improves performance and reliability while giving finer control over storage drives. Most importantly, this also allows us at Seagate the best possible vertically integrated design so that new capabilities introduced at the device level can be immediately leveraged by the object storage system without requiring us to wait for some third-party file system to add support for new device capabilities In addition, ongoing device innovations like HAMR and multi-actuator are being developed in parallel with CORTX to ensure that their capabilities and efficiencies are delivered to end users and applications with the quickest possible path to productization. Finally, CORTX has an integrated scale-out, auto-indexed key-value store (KVS) to simultaneously enable scalable labeling and search of zettabytes of data. Just saving the at-risk data is not enough—we must simultaneously enable fast search across these massive data sets. CORTX has been designed for extremely optimized metadata search, as well.

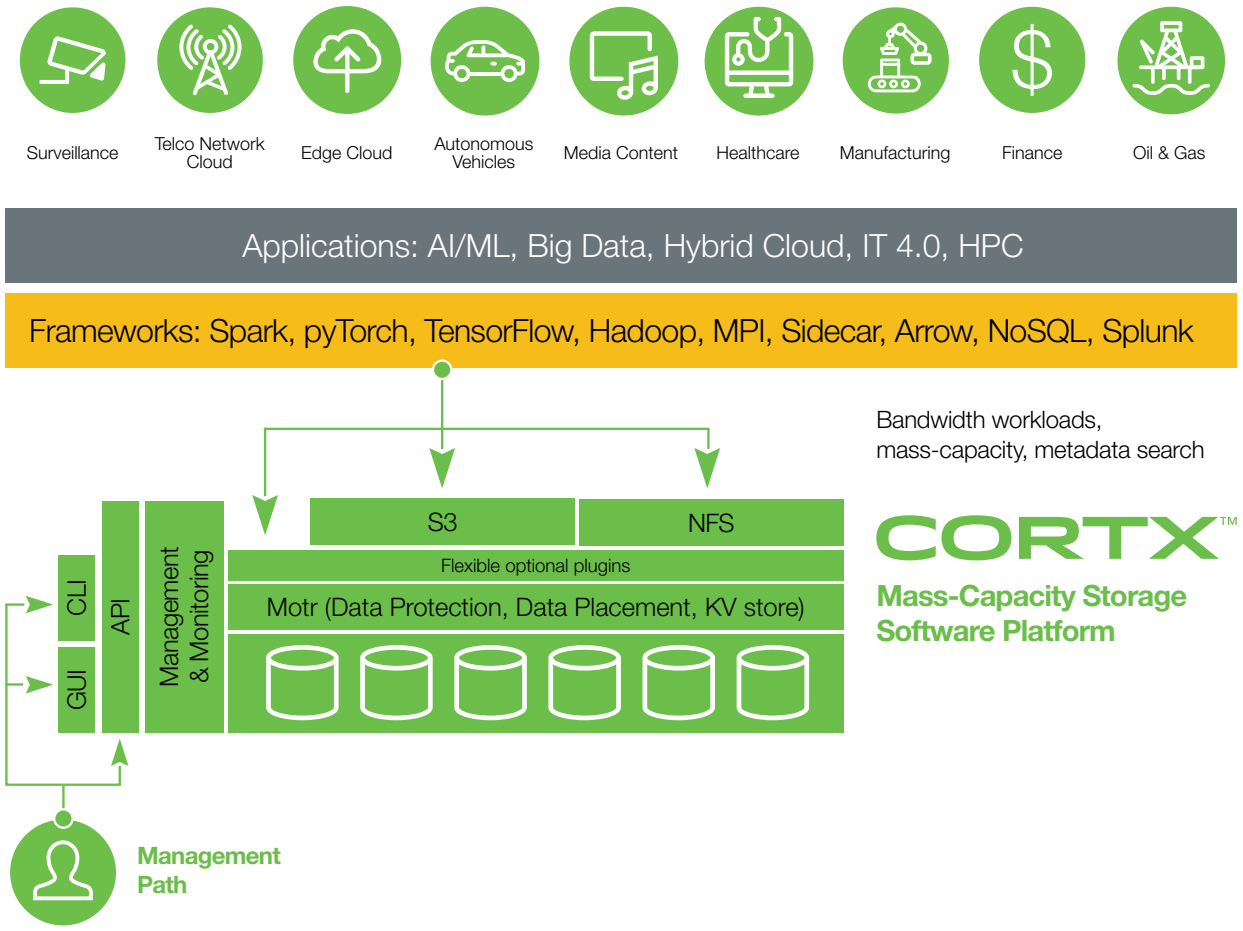
The diagram below shows CORTX's

2 https://www.seagate.com/solutions/mach-2-multi-actuator-hard-drive/

3 https://www.seagate.com/our-story/introducing-hamr-technology

intended role in data centers servicing a wide range of targeted segments via a host of frameworks and protocols. CORTX is uniquely optimized for mass-capacity hardware devices and supports bandwidth workloads, capacity workloads, and metadata search. It is an object storage platform specially designed to support the largest HDDs.

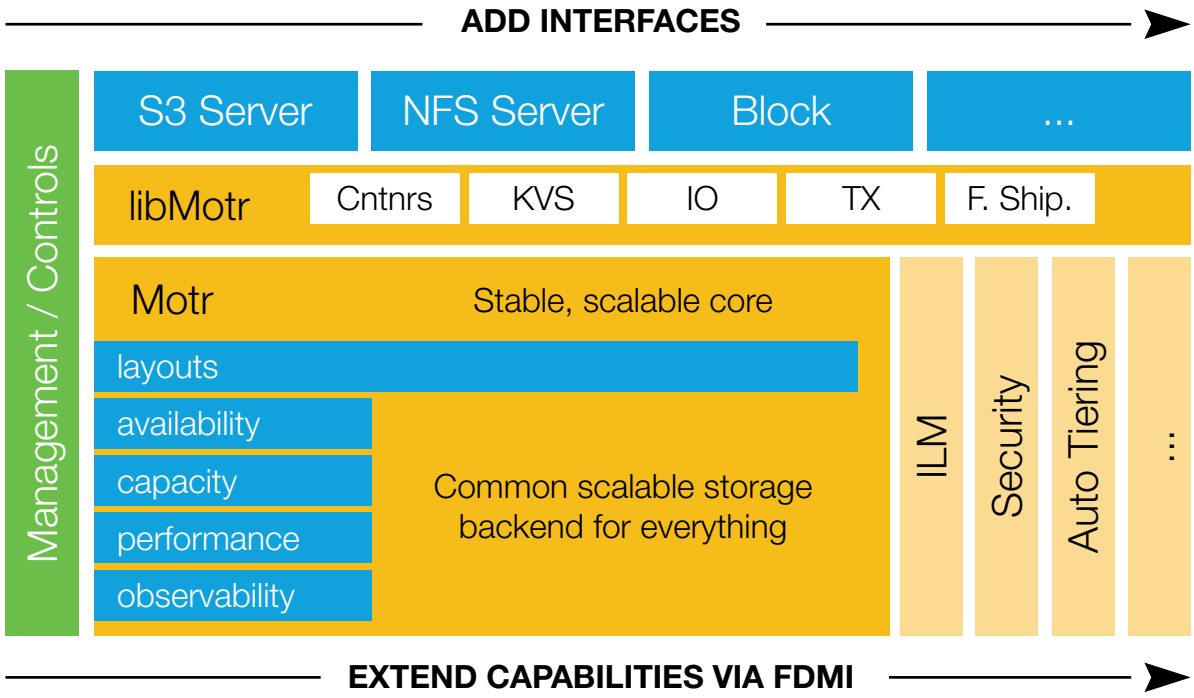
FIGURE 2 The role of CORTX in the IT 4.0 ecosystem



CORTX Extensibility

The diagram below shows the high-level architecture and highlights the extensibility features of CORTX. Developers can easily add additional protocols and interfaces by writing modular interfaces to libMotr to complement the S3 Server and NFS Server interfaces. CORTX also provides a simple interface by which core modularity can be added into the data path. Using a low-level interface called File Data Manipulation Interface (FDMI), developers can add modules via a publish-subscribe mechanism to intercept and augment, or even transform, the CORTX data path. As is shown in this figure, FDMI can be used to add Information Lifecycle Management, security, tiering, and more.

FIGURE 3



At the core of COTRX is the Motr module, which provides data and metadata storage. Motr uses an internal, RAM-based KVS to store object metadata—i.e., data layouts, access control information, security features, etc.—in a compact, efficient form. Motr also provides a user-facing application programming interface (API) for KVS, enabling storage systems users to use KVS for arbitrary small metadata and to provide the fastest access possible to this type of data. For example, the NFS Server module uses Motr KVS for storing abstractions like i-nodes, and user applications can use the Motr KVS to store labels about their data to enable subsequent search.

About the CORTX Community

The CORTX Community strives to create the world's best mass-capacity scalable object storage system. To reach our goals, the community aims to embody four core principles:

Inclusive—Our ambitions are global. The CORTX community is too. The perspectives and skills necessary to achieve our goals are wide and varied; we believe in creating a community and a project that is inclusive, accessible, and welcoming to everyone.

Open—We are dedicated to remaining open and transparent. We believe in keeping CORTX Community code freely and fully available to be viewed, modified, and used without vendor lock-in or other in-built limitations.

Inspired—CORTX is all about the challenge. Our goals are not small: we want to build the world's best scalable mass-capacity object storage system, one that can work with any hardware and interoperate with all workloads. CORTX is built on hard work, ingenuity, and an engineering mindset. We embrace hard problems and find inspired solutions.

Evolving—CORTX is continuously growing and adapting. As a community project, there is no limit to its development. We continuously make room for improvement and welcome the opportunities offered by the ever-evolving nature of community projects.

Ready to join us?

The CORTX Community welcomes everyone. There are many ways to participate, but all of them start with a visit to the CORTX main repository at **https://github.com/Seagate/cortx** or by contacting us at **cortx-questions@seagate.com**

CORTX is an open, evolving project with ambitious goals. Building the best scalable mass-capacity object storage system cannot be done in a vacuum by one company or one team. Building an adaptable, flexible solution that works under more than the hothouse conditions means making room for anyone and everyone who has something to contribute. If you are interested in what we're trying to accomplish and intrigued by hard problems, here are a few ways you can get involved:

Clone, build, test. Put CORTX through its paces and make sure it is truly functioning across as many systems as we want it to.

Coding—help CORTX grow by contributing to the code and helping it reach its full potential.

Documentation—read through the documentation (perhaps while building and testing) and discover where it falls short, or where expectation doesn't meet reality, so that everyone who needs to use it can be successful.

SEAGATE