Komprise Intelligent Data Management

Architecture Overview

[White Paper]

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Explosive data growth requires a re-think of how data is managed. Storage capacity is running out, backups are taking longer, and budgets can't keep up with the unstructured data deluge. Environments are getting more complex as enterprises are shifting to a hybrid, multi-cloud strategy. Managing data within vendor silos leads to poor visibility, proprietary lock-in, and ballooning costs. Komprise provides a standards-based, modern data management solution architected to put you in control of your data with unprecedented simplicity – by giving you visibility into all your data, moving data to the right place at the right time efficiently, and providing native access to data at every tier without proprietary lock-in.

Komprise Intelligent Data Management has been designed from the ground-up to simplify and put you in control of data, no matter where data lives. Our analytics-driven approach works across File and Object storage, across cloud and on-premises, and across storage and backup architectures to give you a single consistent way to manage data. Get instant insight into all of your unstructured data—wherever it resides. See patterns, make decisions, make moves, and save money—all without compromising user access to any data.

Komprise puts you in control of your data while simplifying data management by creating a lightweight management plane across all your data silos without getting in the path of data access. This white paper describes the key components of the Komprise architecture and how Komprise overcomes the limitations of legacy and proprietary data tiering and data migration solutions.

By 2025:

- 175 ZB of data
- 90% unstructured

—IDC[1]

We see immense benefits of using Komprise as it provides a ubiquitous layer across storage to analyze, manage and migrate data.

Jay Smestad
Senior Director Infrastructure Architecture
Pacific Biosciences

[1] 2019 IDC Forecast for 2025. IDC.com

TODAY'S DATA MANAGEMENT CHALLENGES

Massive Growth Within Flat Budgets

Data is growing fast – nearly 90% of the world's data was created in the last two years and enterprise data is doubling every two years. The challenge is how to retain all this data (as much of it is valuable) while keeping within flat budgets. Enterprise IT organizations need to do more with less by using the right mix of cloud and on-premises file and object storage options.

Right Data, Right Place, Right Time

Over 60 to 90% of enterprise data is cold and infrequently accessed within months of creation, but is often stored, backed-up, replicated and managed in the same way as active data. This is because there have been no easy approaches to systematically identify, move and access data without tedious manual processes and disrupting users. Cold data costs are even higher with the cloud given the ongoing OpEx costs of storing and retrieving data from the wrong cloud tier. A systematic way to continuously understand data usage and dynamically move the right data to the right place at the right time without impeding user access is required. Legacy data management solutions are too costly and complex, and cloud tiering solutions provided by storage vendors are too limited, proprietary and create unexpected cost overruns.

Legacy Data Management Challenges:

- **Costly –** Require expensive enterprise licenses and upfront infrastructure investments.
- **Complex –** Multiple moving parts such as storage agents, hardware, software, and databases to manage.
- **Brittle –** Static stubs can be corrupted or orphaned, agents need to be kept up to date as the storage evolves, and detailed rules need to be specified and managed.
- **Disruptive –** Performance slowdowns due to the management overhead they generate, and user disruption by not maintaining transparent access to moved data.

Cloud Tiering and Storage Pools/Tiering Challenges:

- **Limited –** Intelligent tiering and cloud tiering solutions are usually limited to a few simplistic policies (e.g. any data not used in 60 days) and limited to a few storage choices. Enterprises need a robust data management solution with flexible policies to address different groups and their unique needs.
- **Proprietary –** Tiering solutions move data at the block level and so the data cannot be directly used from a lower-cost tier such as the cloud. The cloud is used as a cheap under-the-hood storage tier, meaning that cold data cannot be directly accessed in the cloud nor can it be used with native tools in the cloud or 3rd party applications for AI, ML, or compliance use cases.
- **Cost Overruns –** Cloud tiering and storage tiering or "Pool" solutions have unexpected costs since they are proprietary and access typically causes rehydration of the cold data and results in expensive egress costs from the cloud.
- **Difficult to Manage –** Although storage vendor-based tiering may sound like a simple choice, they are not suited for tiering to the cloud and will lead to performance degradations of the file system. These solutions are limited in the storage they support, create proprietary silos, do not provide visibility into the data that is being tiered, and they lock you in. Switching to another vendor in the future is not simple.

PRINCIPLES OF KOMPRISE TECHNOLOGY

Komprise is Built Upon the Following Design Principles:

Simple
Komprise is simple to deploy and operate — it requires no proprietary interfaces or complex infrastructure setup.

Open
Komprise works using open standards — NFS, SMB/CIFS and REST/S3 — without the use of proprietary stub files or agents.

Vendor Agnostic
Komprise is built on open standards and works with any storage supporting those common standards — allowing you to keep your preferred vendors and manage data seamlessly across multiple vendor storage devices. Komprise future-proofs you and allows you to switch vendors any time.

Analytics-Driven
Komprise uses the analytics from your data usage and growth patterns to provide an ROI-driven approach to optimally manage your data based on your unique data needs.

Transparent
Komprise moves data transparently which means it is fully accessible from the source as files, exactly as before, and the data is accessible as files or objects from the target. This ensures users are not disrupted and can still find the cold data where it was originally located on the source. Komprise maintains native access to the data on the target, so you are always in control of your data no matter where it resides.

Elastic Scaling
Komprise scales elastically on-demand — there are no central bottlenecks, databases or servers to limit scalability.

No Lock-In
Data is always accessible from your source storage and your target storage, even if Komprise is taken offline.

Data Management-as-a-Service
Komprise does not require dedicated hardware or upfront infrastructure investments. Komprise runs as a hybrid cloud service or as a fully managed data management-as-a-service (DMaaS) in the cloud.

Non-Intrusive
Komprise analyzes and manages data in the background, with no impact to storage and network performance, and outside the hot data and metadata paths.

Adaptive
Komprise throttles back as needed when your storage or network are in active use, so you never have to monitor or schedule when Komprise runs.

KOMPRISE ARCHITECTURE

The 6 Key Components of the Komprise Architecture:

1 Komprise Elastic Grid

Komprise handles any scale of data using an elastic scale-out architecture that has no central bottlenecks. Start with a Komprise Observer virtual machine and simply scale by adding more. Komprise automatically load balances and creates fault-tolerance across the elastic grid. There are no central databases or bottlenecks to limit scalability. Komprise scales to handle billions of files and tens of petabytes of data with a lightweight distributed architecture. There is no single point of failure – Komprise uses fault-tolerant architecture principles to create a resilient grid with Observer virtual instances that provide failover to one another in an active-active configuration without requiring any dedicated infrastructure.

2 Komprise Policy Engine

Komprise uses high-level policies that drive its automated data management and data movement. Unlike rule-based architectures that require you to specify low-level rules for every condition and can create rule conflicts and unnecessary complexity, Komprise provides a declarative model where you specify high level policies such as when data is considered cold or how you want data replicated. Komprise does the rest – you can set the same policy across multi-vendor clouds and storage, making it easy to manage data at scale across an enterprise that has a heterogeneous storage unit.

3 Komprise Transparent Move Technology™ (TMT)

Komprise, with its patented Transparent Move Technology (TMT), moves data transparently using standard file system constructs. In order to provide a storage and cloud agnostic architecture, Komprise avoids the use of static stubs and agents. It archives files such that the archived files continue to be accessed from their original location as files, while the data actually resides as objects in the cloud. Komprise provides this file-to-object translation without requiring re-hydration back to the source. Files archived by Komprise are also accessible as native objects from the cloud without going through Komprise or the original file storage, so there is no lock-in. Finally, Komprise does not get in front of hot data access.

4 Komprise Cloud File System

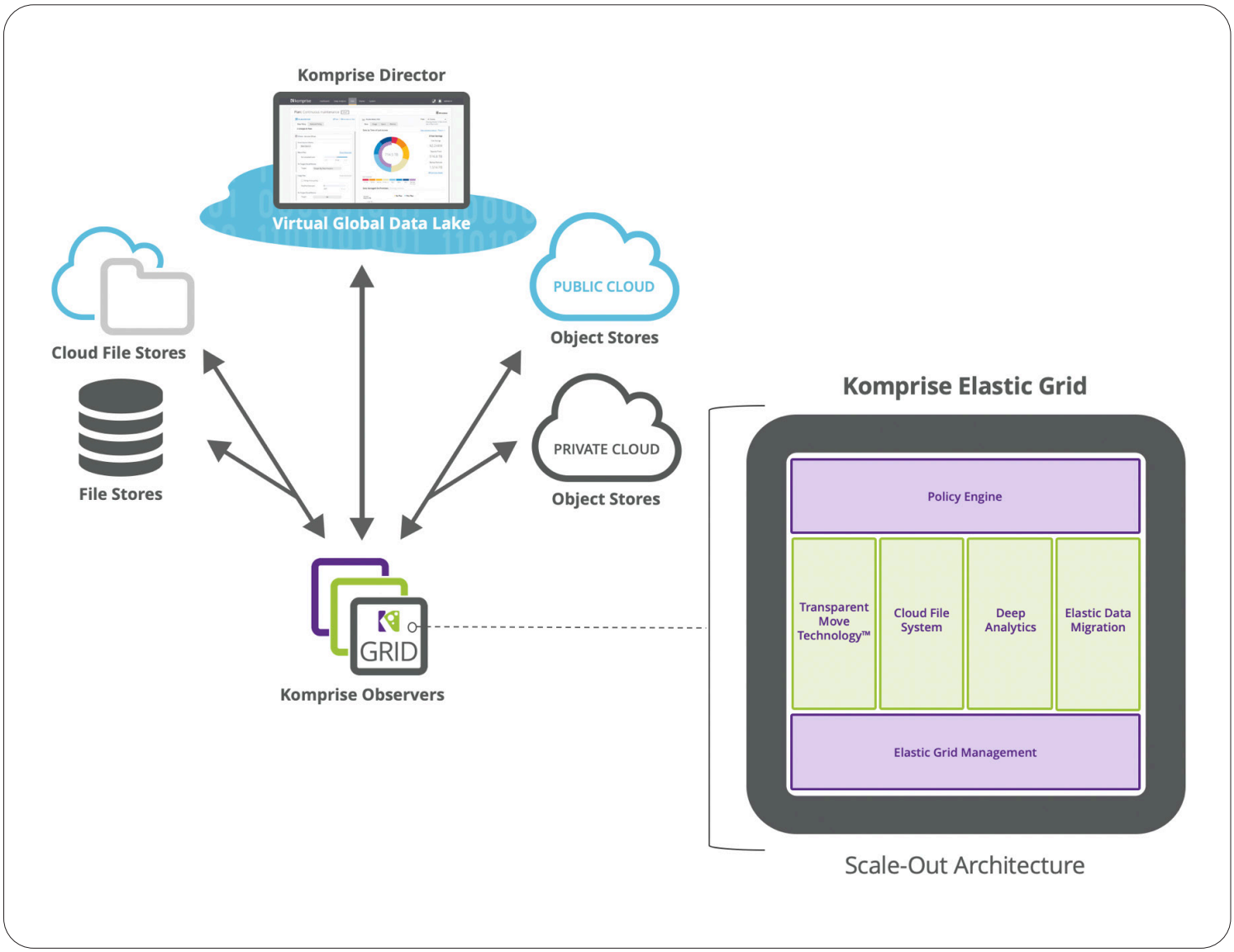
Komprise leverages the sources and targets themselves to create a redundant, overarching namespace. When Komprise moves data, it keeps the data object in native form at the destination and also preserves all the file metadata and access controls. When an archived file is accessed from the source, the Komprise Cloud File System receives the access request and translates the cloud object into a file without having to rehydrate it back to the original storage. This provides a highly efficient way to conserve space on the original Network Attached Storgae (NAS) by leveraging cheaper cloud object storage classes, such as Amazon S3. It also ensures the archived data in the cloud can be natively accessed as objects in the cloud without using Komprise or the original storage, so there is no lock-in. This is very important because it allows you to use native cloud tools to access, process and extract value from your archived data. Finally, the Komprise Cloud File System ensures a file view of the archived objects, independent from the source, so data can be directly used in the cloud either as files or as objects.

5 Komprise Deep Analytics

Komprise automatically creates a virtual global data lake by indexing metadata across all your on-premises and cloud storage. Now, you have one go-to place to search, tag, and operate on your data across your enterprise. In essence, this approach addresses mass data fragmentation by automatically creating one “logical silo” without restricting you to a single storage platform. This not only enables IT teams to gain deeper insights and greater control over data, but is also used to optimize and drive how Komprise moves, archives, replicates, and manages data to suit your particular environment and unique data needs.

6 Komprise Elastic Data Migration

Komprise eliminates the cost and complexity of data migrations with **Komprise Elastic Data Migration**, which analyzes and automatically parallelizes data movement at every level to maximize performance. Elastic Data Migration migrates up to 27 times faster than generic tools across heterogeneous cloud and storage. You can then setup hundreds of migrations and run them faster at a fraction of the cost. Komprise ensures full file fidelity with MD5 checksums on every file.



Komprise Elastic Grid Scale-Out Architecture, including the key architectural components of Komprise Intelligent Data Management.

Organizations working with Komprise have reported improvement in both TCO and ROI after adoption. In general, the more storage capacity under management, the bigger the savings.

GIGAOM
2020 GigaOM Radar for Unstructured Data Management

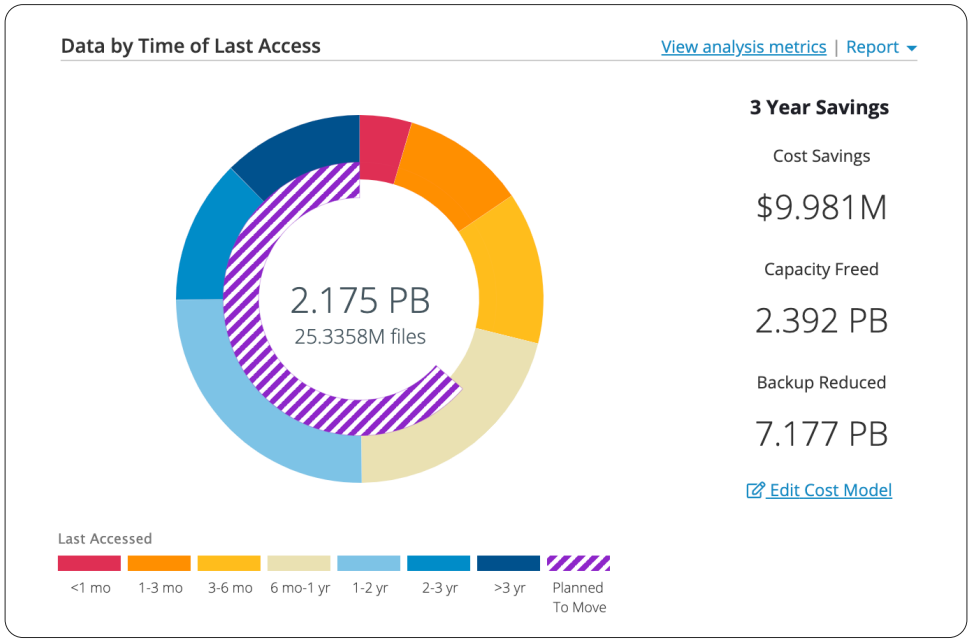
HOW IT WORKS

Dynamic Data Analytics

Deploying Komprise is simple — download the Komprise Observer and point it at existing storage via NFS or SMB/CIFS or S3. Within fifteen minutes, Komprise Dynamic Data Analytics enables you to:

- Analyze data across all your storage to understand who, what and when
- Show trends to determine how fast your data is growing
- Interactively forecast savings based on your policies

Storage and backup vendors that offer analytics require your data to be on their hardware first before they'll give you these insights. This forces you to make costly investment decisions and move or copy data before you've had a chance to first understand it. (If you could, you probably wouldn't make the same decisions.) Komprise lets you know first, then move smart and take control by providing analytics in place, across your hybrid, multi-cloud storage environments.



Easily estimate cost savings of different data storage decisions.

Here are three ways Komprise Dynamic Data Analytics helps manage unstructured data:

1. **Analyze all your data before investing, copying, or moving**

   Komprise lets you analyze data across your storage environments without first requiring a move or copy. We do this by connecting to your storage and analyzing your data via standard protocols, such as NFS, SMB or S3. This allows Komprise to analyze your data in-place without needing to import it into a proprietary format. No proprietary interfaces, agents, or clients on your storage—just a standards-based approach that gives you critical insights into your data—wherever it resides—allowing you to make the best, most-informed decisions.

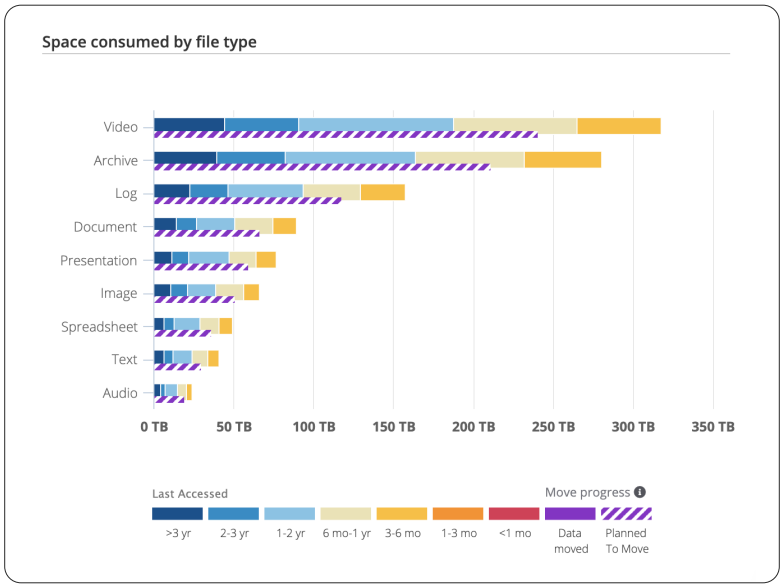
2. Analyze usage metrics in hours, not weeks

Most unstructured data analytics solutions take weeks to crawl and index all your data before they can provide insights on billions of files and petabytes of data. Komprise delivers Dynamic Data Analytics in hours—even on billions of files—using patented data analytics and aggregation techniques. It simplifies building your data management plan by allowing you to:

- Find files by various criteria such as file types, sizes, owners, top groups, last access time, etc.
- Run "what if" scenarios and get subsequent capacity needs and cost savings in seconds
- Manage and move your data the way you need, to save the most

Run "what-if" data scenarios and get an instant cost analysis.

Want to know what would happen if you moved all data untouched in over a year to the cloud? Komprise provides instant analysis based on your data, your costs, and historical data growth patterns.



See what kind of data is being used how often.

3. Dig deeper across your enterprise-wide virtual data with Deep Analytics

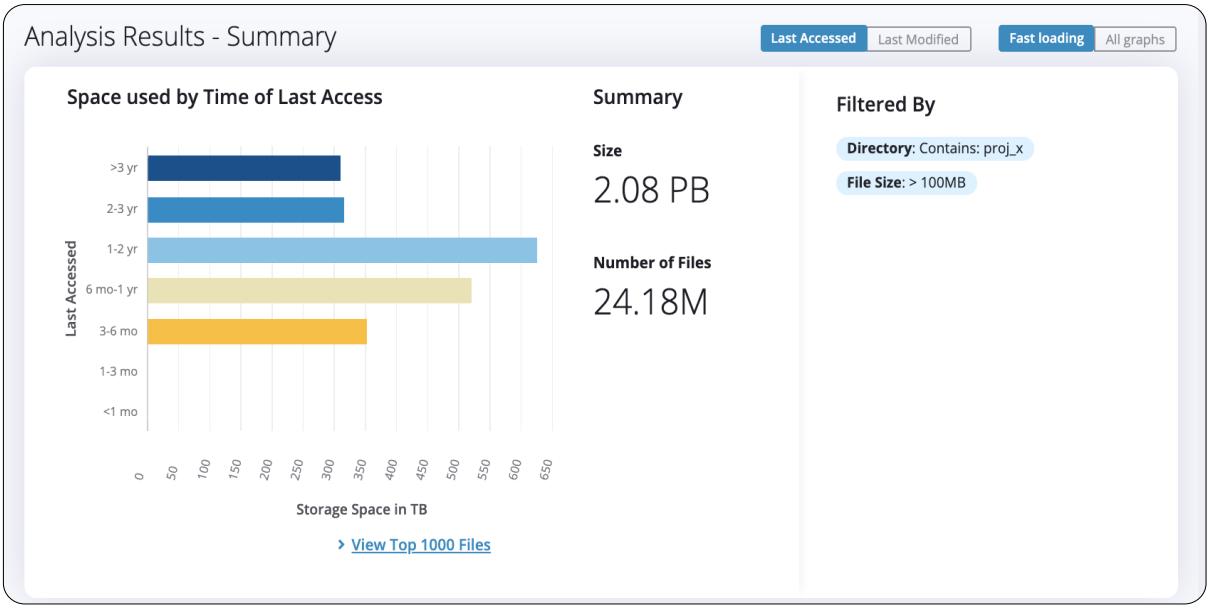
Komprise provides an intuitive way to search and find specific files that fit your exact criteria across all your storage. Simply build your custom queries and Komprise Deep Analytics shows you both summary information and detailed reports on the files that fit your criteria. You can tag the data you find. Using APIs you can also tag files and objects as they are created. This dynamic approach allows you to then run queries based on your tags and build real-time virtual data lakes on the fly, without having to first move the data. You can continually leverage these data lakes for big data, AI, and ML applications. This is available both via a user-interface and an API.

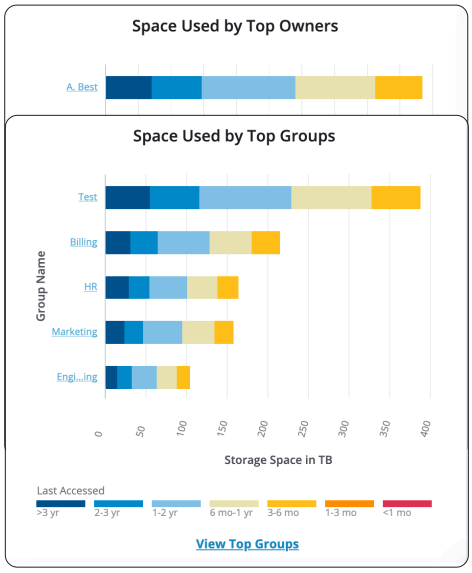
Virtual Global Data Lakes

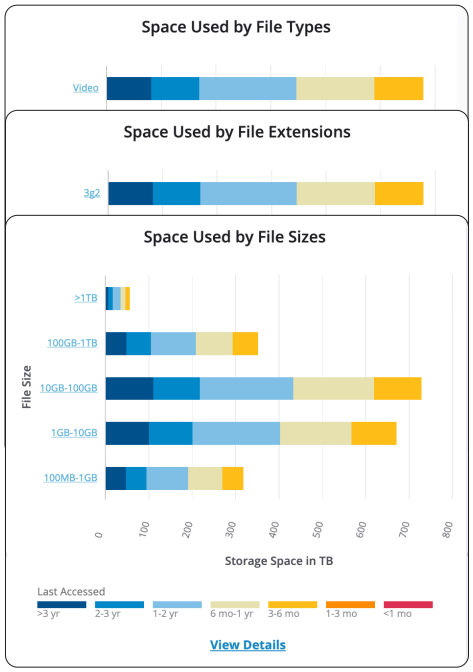
Komprise provides a catalog of **ALL** your data across all data centers and clouds

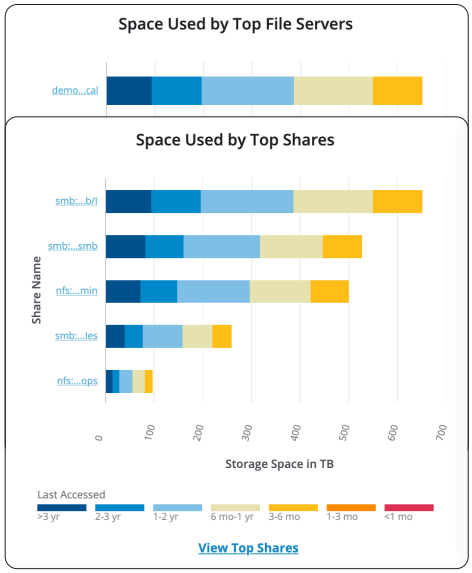
- Global search
- Tagging
- Operations

Komprise Deep Analytics Reports:









Drill down for detailed reports filtering by specific criteria. This example shows the summary and detailed reports of filtered data with file names containing "proj_x" and file sizes greater than 100MB.

Transparent Data Movement

Once Dynamic Data Analytics gives you the insight to make the right decisions, it's time to take action. Komprise Transparent Move Technology (TMT™) works seamlessly with any storage and backup solution, allowing you to make the data moves you want, when you want, without disrupting user access. TMT enables you to create the logical architecture you want without disrupting the physical architecture you have. This gives you the best of both worlds - empowering you with the agility to evolve rapidly without creating user friction and hurting business productivity. For instance, you can move to a cloud architecture without eliminating your file-based NAS and without disrupting user access. Komprise achieves this virtualization without creating any interference to hot data or metadata through the patented TMT technology.

Komprise customers typically discover that **~80% of their data is cold**. TMT lets you archive that data off primary storage so you can:

- Save on storage and backup license costs resulting in as much as 70% cost savings
- Cut your backup time
- Enhance performance/availability of hot data

Data rehydration **eliminates** storage savings.

Issues with proprietary tiering solutions provided by storage vendors:

Some storage vendors provide a proprietary tiering solution for cold data. These solutions don't provide data analytics nor a granular policy that can be customized for each share. If you use such proprietary tiering solutions with a third-party backup solution, all archived data may be rehydrated. You'll need to maintain the full capacity of your existing data to rehydrate all archived data, thereby eliminating any cost savings!

Controlling access to your moved data is part of storage vendors' business model.

TMT works with all standard backup solutions and does not rehydrate the data, thereby ensuring significant cost savings. TMT makes moving cold data to less expensive storage simple. You can literally make your data moves with a click of a button, without any disruption to users, applications, data protection workflows, or access to mission-critical hot data.

The biggest issue with proprietary tiering solutions is that they are implemented at the block level and not at the file level. When they tier data (as blocks) to the cloud, they access and interact with cloud storage using the same complex block management mechanism and protocol designed for their internal storage platters. Unfortunately, the cloud does not lend itself well to this approach. The cloud has much higher access latency and does not provide the ability to write portions of an object needed for block management. As a result, this block-based approach is highly inefficient, leads to unnecessary cloud egress costs, and potentially slows down overall performance of the storage unit. It is for these reasons, storage vendors recommend using their own on-premises object storage rather than the cloud to mitigate some of the issues. Note, Komprise with TMT works at the file level, does not have these issues, and works equally well with cloud storage as well as on-premises object stores.

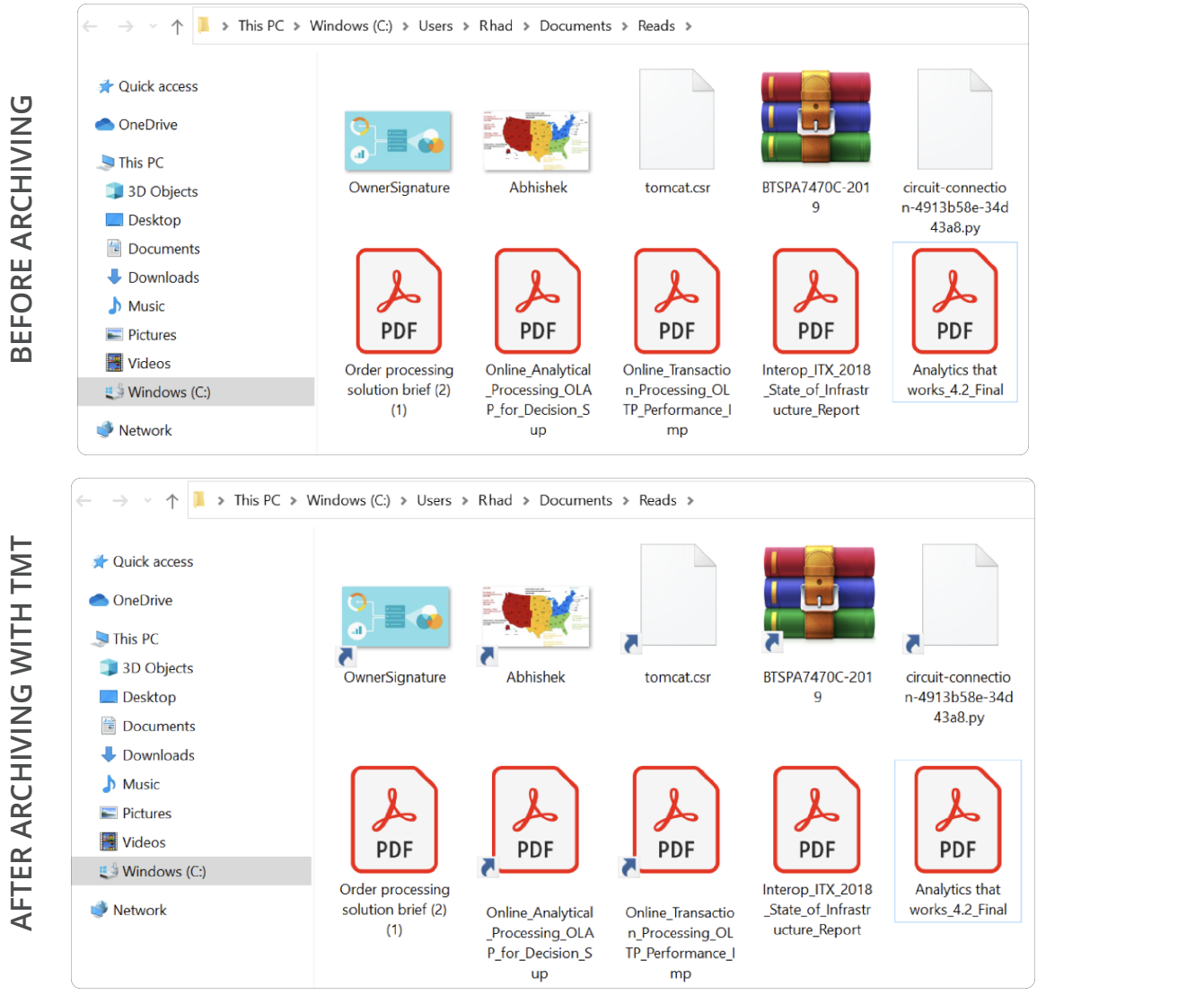
TMT addresses the following key issues critical to archiving cold data:

1. **No disruption to users, apps, and data protection workflows**

   Users shouldn't know that their data's been archived, no matter where it's been moved. They should be able to go where they've always gone—same source, same directory—and access it the same way, even if it's in the cloud.

   Most data management solutions don't provide this kind of data access transparency. End users have to use a separate application or generate an IT support ticket to find their archived data, which creates an unnecessary drag on productivity and a burden on IT.

   Komprise delivers transparent access by using standard protocol constructs when moving data. When you move a file, a symbolic link containing all the properties of the original file is left behind as a pointer. Users and apps continue to see and can open the file from the original location keeping all the permissions and access control intact. No invasive agents or stubs means no disruption to users, applications, or the data protection workflows. This transparency applies to wherever your data resides. TMT is leveraged universally and seamlessly across all storage tiers, including cloud, object storage, and backup.



Users are unaffected, accessing moved data exactly as before. This example shows thumbnails of files before they were archived and after. The only change is the small arrow on the thumbnail.

2. **Works at the file level, not block level, with full metadata fidelity**

Storage vendors use block-level tiering to move data out of the file server and into an object or cloud tier. Block-level tiering was developed for internal use for moving file servers ***within*** blocks between the various tiers to increase performance while reducing costs. Hot blocks and metadata are kept in the higher, faster, and more expensive storage tiers, while cold blocks are migrated to lower, less expensive tiers.

In theory, it makes sense to extend this notion to also transparently archive cold data to the cloud, but unfortunately it does not translate. The cloud is not an internal storage platter with microsecond access latencies. Rather, it is an object store across a WAN. It is not designed for partial writes of small blocks. Instead, it is designed for writing larger objects, which cannot be altered, but must be re-written completely when a change is to be made. As a result, block-level tiering to the cloud degrades performance of the storage unit and increases cloud egress costs. Storage vendors do not recommend large-scale tiering to the cloud.

Since blocks, not files, are written to the cloud, the data written to the cloud is meaningless and cannot be directly accessed from the cloud. Block-level tiering forces users to access moved data ***only*** through the vendor's hardware or software—direct access is no longer an option. Finally, using third-party backup software with block-level tiering may result in rehydrating all the cold data back to the storage unit thus eliminating any cost savings.

File-level tiering maintains full file fidelity and preserves all the attributes and metadata along with the file at each tier.

File-level tiering is a more advanced technology and is standards-based. Tiering at the file level means that the file and all its metadata moves to the new tier. It maintains full file fidelity and preserves all the attributes and metadata along with the file at each tier. Applications that rely on attributes of the file are unaffected. And, users can access data ***directly*** from the target storage and be able to return it to the source storage exactly as they did before.

3. **Zero access interference to hot data or hot metadata**

Many data management solutions sit in front of your primary storage and divert requests for the cold data to another location. In general, these solutions promise some form of data virtualization or metadata offloading. But being in front of your NAS storage impacts the performance of the hot data since it introduces a middleman. A "traffic cop" now directs data access, which is a tremendous risk for all your company's data. A failure in this system creates an access nightmare. Just as if the traffic cop took a break, you'd have a major logjam—and when access to all of your data is lost, you'll hear far worse than honking horns.

Komprise analyzes and manages data in the **background.**

- No impact to storage and network performance
- Outside the hot data and metadata paths

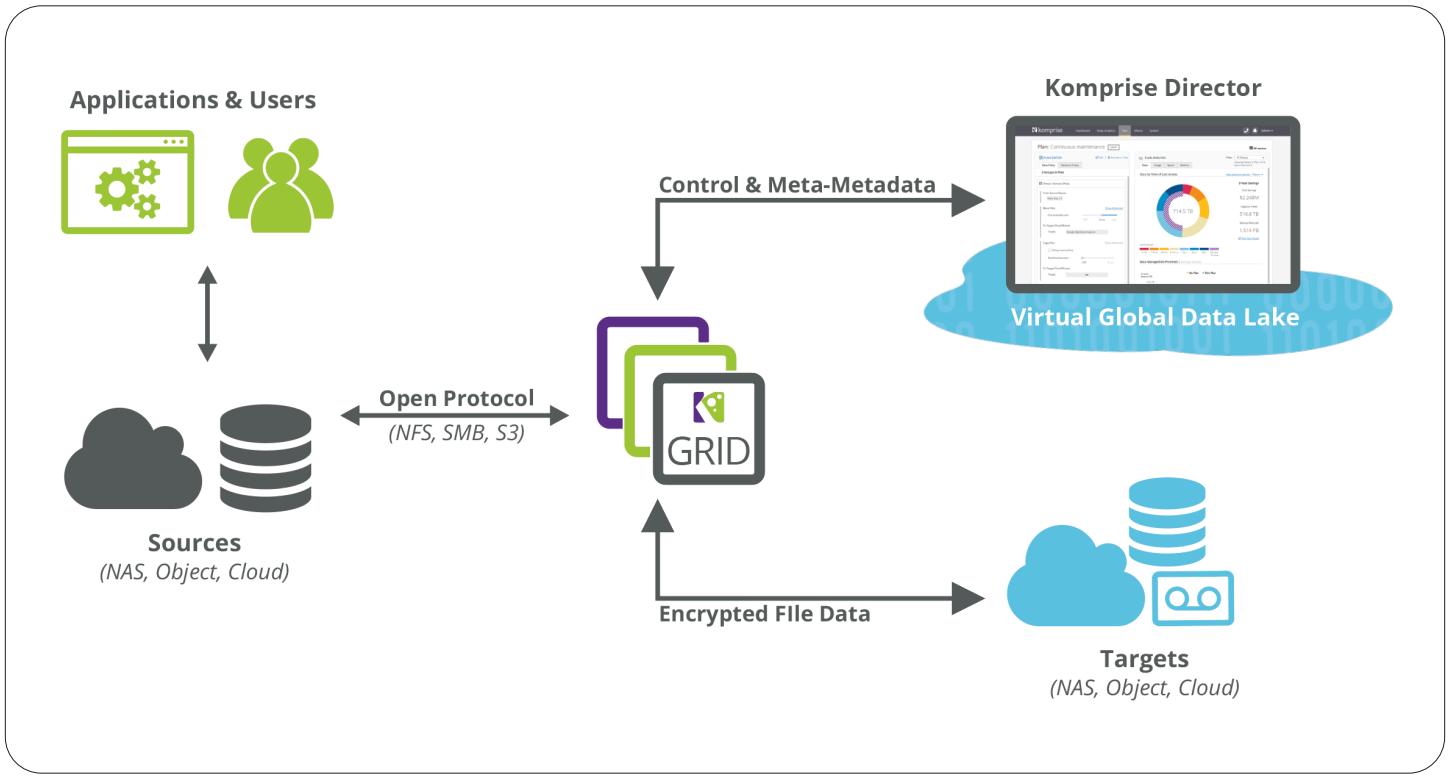
This approach also requires that scaling be based on hot rather than cold data access rates. Since hot data is 99.999% of your data access, this means that the "man-in-the-middle" device must be able to handle the massive hot data traffic requests. As this traffic increases with your data growth, the "man in the middle" must scale accordingly and still handle access spikes. If you don't plan accordingly, you'll decrease performance on your new flash storage. These types of solutions—and their issues—have been around for years losing customers due to high cost, decreased performance, and risk.

Direct Data Access

Puts you, not your vendor, in control of your data.

With the advent of big data analytics, AI, and ML, the ability to index, search, and operate on all your data is a game changer. It's also a fundamental requirement for ever-increasing compliance and legal data-hold requirements. It's important to have direct data access to all your data without going through the original source.

Komprise makes this possible because it transfers your company's NAS data to a cloud or secondary storage target as a file with its complete metadata, whether that target is an on-premises storage (NAS or object storage) or cloud storage. The file is stored in the format native to the target storage device, and the metadata is stored in NFS or SMB format, depending on the protocol used to archive or copy the data.

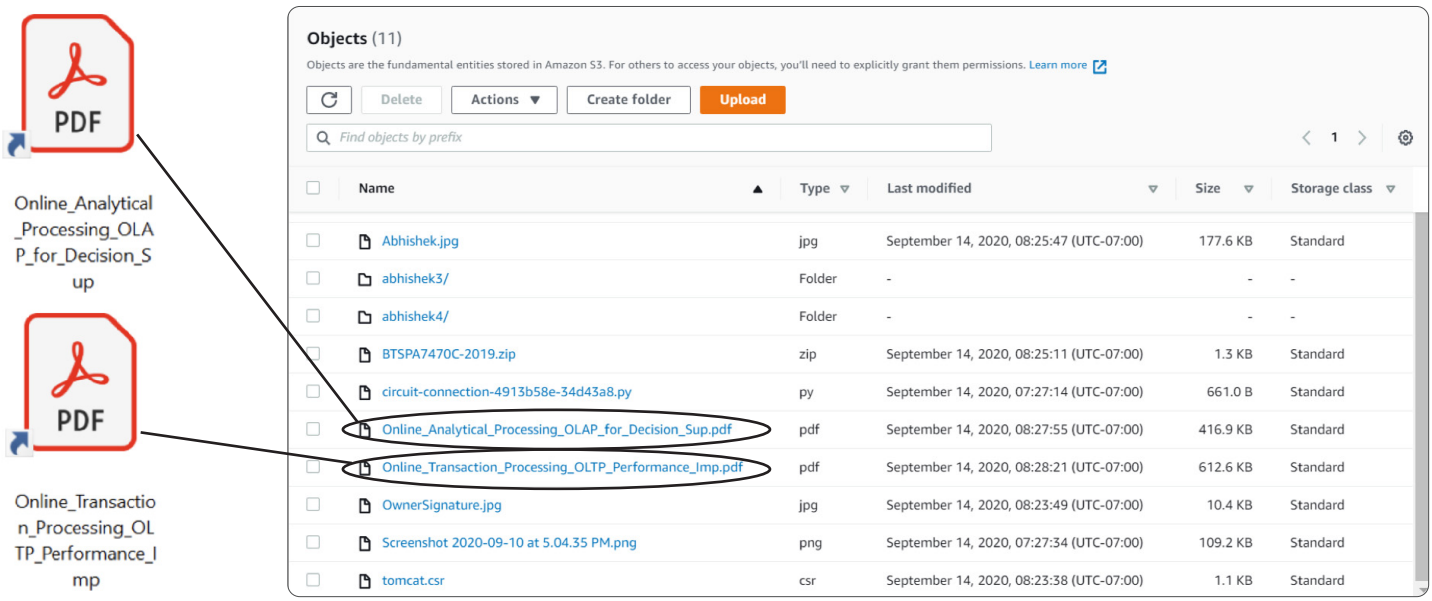


The Komprise scalable architecture works across any storage using standard protocols—with no interference to hot data, metadata, or network paths. Access data with full fidelity from anywhere with no lock-in, keeping you in control of your data.

Direct Data Access Benefits:

1. **Access data from any tier without going to the source**

   When you move your data with Komprise, it can be directly accessed from the target device using its native protocol. Let's say you move data to the cloud using Komprise. You can access that data transparently from the source and directly from the target using standard S3 tools. You can view and access that data as files by mounting the Komprise Cloud File System. This ensures your data doesn't get locked into any one storage or backup solution—or to Komprise for that matter.



The archived files are available as objects in the cloud, as shown, maintaining cloud native access.

2. Access data in native format

At Komprise, we believe you should own your data and not be locked into a proprietary solution. When we archive or copy your data, we put it in a form that's native to that storage system. This ensures that archived data is not locked away and can be continually used to extract value. You can use third-party applications, cloud native tools, or Komprise applications to extract that value—Komprise never locks you in. We preserve all the standard metadata and the extended metadata, such as tags with the data wherever it moves, so your files retain their full context and remain usable wherever they go.

3. No rehydration necessary

Block-level tiering, used by many data management solutions, requires rehydrating archived data before it can be used, migrated, or backed up. This approach negates much of the benefit of data archiving in the first place. For instance, if you archive 75% of your data but must rehydrate it when backing up, you've saved **nothing**. Or, if you want to end-of-life your 1PB storage system from which you've archived 3PB of data over its lifetime, you'll need to rehydrate all 3PB before you can migrate off that system. Komprise file-based tiering eliminates these rehydration issues:

- Symbolic links left on the source when Komprise archives files are understood by other applications.
- Backup software will backup the symbolic links without rehydrating the files they point to.
- Restores will restore the links, which still point to the same files, so third-party backup applications will function without rehydration.
- When migrating data from the source, the symbolic links get migrated without getting rehydrated.

By using industry-standard constructs, Komprise seamlessly operates with your storage and backup solutions and other applications without requiring any customization.

SUMMARY

Komprise is an analytics-driven data management strategy that allows you to know your data before making decisions about it. The Komprise architecture is designed for today's massive scale of data to work across multi-vendor clouds and storage with a lightweight non-intrusive data management solution. Komprise is dedicated to data autonomy, allowing data owners to analyze, manage, and move their own data, free from proprietary technologies. We allow customers to transcend storage silos, storage vendors, and storage technologies, providing the flexibility you need to scale as your data grows—wherever it goes—and to be agile and responsive as your needs change. With Komprise Intelligent Data Management, you are able to know first, move smart, and take control of massive unstructured data growth while cutting 70% of enterprise storage, backup, and cloud costs.

> *For enterprises that have been reticent about moving file content to cloud storage because of a lack of data management, Komprise can assist by identifying the business value (or lack thereof) of content before migrating it to IaaS environments.*
>
> **Gartner**

ABOUT KOMPRISE

Komprise empowers businesses to take control of their data and save costs—with no interference to applications, users, or hot data. Komprise Intelligent Data Management is the foundation for analytics-driven data management which is key to putting data in the right place at the right time across all storage. Analyze, move, and easily find your data with Komprise.

LEARN MORE

To learn more about Komprise Intelligent Data Management visit **Komprise.com/product**

komprise

Komprise, Inc.
1901 S. Bascom Ave. Suite 500
Campbell, CA 95008
United States

For more information:
Call: 1-888-995-0290
Email: **info@komprise.com**
Visit: **komprise.com**

For media requests email
marketing@komprise.com