Igneous

WHITEPAPER

The True Cost of Cold Data

Igneous
2401 Fourth Ave., Suite 200
Seattle, WA 98121
(844) IGNEOUS
igneous.io

Table of contents

Introduction

For decades, IT administrators and data owners have been in conflict over how to decide whether data is in active use or if it can be archived. IT administrators, charged with managing primary storage capacity and backup SLAs, have been asking data owners to designate data that can be moved off of primary storage and over to cheaper archive tiers.

However, this is at odds with the needs of data owners, who would rather have immediate access to all of their data, just in case – regardless of whether they actually need it or not. This is frequently a learned behavior, as previous archive strategies – such as tape drives going into a salt mine – might as well have been black holes, where their data would never be seen again.

Archive options have improved dramatically over those same decades, with many solutions offering simple restore workflows to make data available in minutes or hours, rather than days or weeks. Some solutions – such as Igneous DataProtect™ – even provide end users with direct access to archived data, effectively eliminating the restore window. Unfortunately, these improvements have so far not translated into more data archived: IT administrators are still largely unable to convince data owners to move their data to cheaper storage.

In order to tip this ongoing conflict toward a resolution, IT administrators need to be able to reliably and regularly identify datasets that are no longer in use, and to show that moving those datasets to an archive tier can save the organization substantial money. That second factor – saving the organization substantial money – is both the most critical, and the most difficult to quantify. How much money can be saved? Where will those savings come from? Does the money saved justify the hassle?

The practical barriers to archive are well known:

- Individual and organizational resistance to moving data from where data owners expect to find it (and often cutting data owners out of any direct access to their data)
- Administrative overhead of fielding requests for user access to archived data
- Slower performance from the archive storage tier

Without accurately calculating the financial incentives, it's impossible to overcome these barriers and justify to data owners why their data needs to be archived.

The real cost of keeping cold data on primary storage

If primary tier NAS capacity was purchased as a capital expenditure, many people will assume that the marginal cost of keeping data on the primary tier is free. However, this is demonstrably wrong. Every terabyte removed from existing primary storage and the associated backup environment generates real savings by delaying the substantial cost of adding additional primary and secondary storage.

Imagine a scenario where a company is starting a project that will require 500 terabytes of primary capacity. If that company doesn't have 500 terabytes of available space, they will either need to add new capacity, or free up space on existing capacity. To add 500 terabytes of storage to the primary tier, connect it to the network, protect it, power it, cool it, and house it in the datacenter can easily cost more than $1M over the next five years. This paper will show that that same organization could instead archive 500 terabytes of unused data from their existing primary storage tier to a low-cost archive tier at a cost of $400K or less over those same five years, gaining the same amount of unused primary storage and saving $600K or more – that's $120K per year, or $10K per month. These significant cost savings could be reallocated to replace aging primary NAS capacity with high-performance all-flash storage, or to accomplish other high-value IT projects that otherwise couldn't be funded.

To get a clear picture of the costs of keeping data in place once it has aged out of active use, we need to add up the full cost of the primary storage that is occupied by cold data. This means including not just the cost of the primary-tier hardware, but also the licensing and maintenance costs, network costs, energy costs, and datacenter costs.

Additionally, nearly all organizations today protect their primary storage with a backup solution. When data is removed from primary storage, there is less data on the primary tier to backup, meaning less backup hardware, software, energy, and datacenter costs as well.

Let's go to the numbers

The following estimates are based on a mixture of publicly available information and firsthand information we have received from real customers that vary in industry as well as scale. We have done extensive research to get a clear, data-driven view of how much organizations typically spend on every aspect of their unstructured data storage environment.

Calculating the total cost of primary storage requires knowing the total cost of hardware, software, energy for power and cooling, and the cost of the datacenter floor and/or rack space.

If you add all of these costs together over the expected life of a primary NAS server, you can aggregate them into a platform-specific cost per terabyte per year. A standardized number such as this simplifies the process of tracking and managing primary tier NAS operating costs.

Hardware and software costs vary by storage vendor, storage (or node) type, and software add-ons that increase functionality, and may even differ significantly from one platform to the next within the same organization. High-performance, all-flash storage tiers will be significantly more expensive per terabyte than standard or more dense primary tiers. For most enterprises, the price will range from **$175-$500 per terabyte per year**, depending on the hardware and software they choose. There are unstructured storage solutions that are less expensive, but they do not provide the performance and functionality typically required for production workloads.

Power and cooling costs will also vary from one platform to the next, with the typical total energy requirements ranging between 30-50 kWh per terabyte per year. Average datacenter electricity prices in the United States are around $0.12 per kWh, giving a total energy cost range of **$4-$6 per terabyte per year**.

Datacenter costs can sometimes appear to be free if the organization owns the building anyway. It's easy to think of this as a sunk cost, but IT teams will still need to account for the datacenter costs as long as space in the datacenter is finite: a rack unit used to host primary storage means one less rack unit available for anything else. Once all rack units are occupied, organizations must either build a new datacenter or rent space in someone else's datacenter.

Fortunately, this means that estimating the cost of a rack unit is straightforward, whether the datacenter already has a cost per rack unit established or not. Typical market rates in the United States per rack unit are between $240-$300 per year. A single 4U primary storage node can host anywhere from 85 to 480 terabytes. That means the datacenter costs per terabyte are in the range of **$2-$13 per terabyte per year**.

Finally, there are **personnel costs** required for managing any hardware or software component within an enterprise. Even if we assume the salary (and benefits) of an IT administrator to be a fixed cost, there is still an opportunity cost that needs to be factored in. Every hour spent by IT administrators racking gear, running software updates, triaging and resolving service incidents, contacting support, managing restores, monitoring system health, and a variety of other tasks is one less hour that could have been spent on other, potentially more high-value IT tasks. Our research shows that the best way to estimate this cost is based off the amount of space any service occupies in a datacenter. Our research indicates that the fully-burdened[1] cost to an enterprise for an IT administrator is around $120,000 per year, and that for managing primary storage, their time will typically cost $375-$500 per rack unit. Using our earlier figure of a single 4U primary storage node hosting anywhere from 85 to 480 terabytes, this gives us an estimated personnel cost of **$3-$20 per terabyte per year** depending on the density of the primary system.

1 "Fully burdened" refers to the total annual cost of the employee, and includes salary and benefits, federal/state/local payroll taxes, unemployment and other work-related insurance premiums, along with equipment, licensing, and training costs.

Adding it up: Primary Storage

By aggregating all these factors together – hardware, software, maintenance, facilities costs, and staffing costs – we see that the total cost of primary storage – referring only to the costs associated with the capacity – has a range of **$184-$539 per terabyte per year**.

Here is an example, of how much it will cost an organization to store their data on their primary tier over a period of several years. We used numbers on the low end of the above ranges, with 1 petabyte of primary at the start of year 1, and growing at a rate of 20% per year.

Year	1	2	3	4	5
Total data on primary	1,229	1,475	1,769	2,123	2,548
Cost of primary hardware + software	$215,040	$258,048	$309,658	$371,589	$445,907
Cost of primary power & cooling	$4,719	$5,662	$6,795	$8,154	$9,784
Cost of primary rack space	$2,458	$2,949	$3,539	$4,247	$5,096
Cost of primary storage FTE	$3,840	$4,608	$5,530	$6,636	$7,963
Total cost of primary	$226,056	$271,267	$325,521	$390,625	$468,750

However, this is not the full extent of the costs of storing data on a primary tier. Data on a primary storage tier is almost always backed up to a secondary location. That means that for every terabyte of primary data, there's an associated cost to backup that terabyte, even if the data is rarely or never used. To get the true cost of cold data, the cost of the backup process and storage must also be accounted for.

The cost of backup and secondary storage

Until recently, backup solutions for enterprises typically fall into two strategies: tape or disk-to-disk (D2D) replication. We have previously discussed why Igneous DataProtect is a superior backup strategy to either tape or replication. For the purposes of this discussion, let's just focus on the costs of the old methods of backup.

First, let's talk about the costs of tape

While lots of organizations look at the unit cost of a tape drive and assume it's the cheapest option available, the truth is often the opposite. As we've just demonstrated, the unit cost of the storage itself is only one part of the equation. Using tape, in particular, for a backup strategy involves a fair number of moving parts. Tape workflows typically require a full (level-0) backup to be completed once per week or once per month, depending on the organization's internal SLAs with data owners. Then, most enterprises will maintain backup versions for some amount of time (often a year or more), meaning that for every terabyte of primary data, organizations have to move at least 12 terabytes or 52 terabytes to tape every year, and are also storing 12-52 terabytes of primary data on tape for every terabyte of data that has been on their primary tier over the previous year. For this reason, we have to look at terabytes moved, not just terabytes stored on the primary tier.

The tape itself might be cheap, but to actually write backup data to tape requires tape libraries, servers, and software, the cost of which can add up very quickly. The easiest way to conceptualize the total **hardware and software cost** of backup is to simplify it to a cost per unit of data moved.

To calculate this number for any specific datacenter can be somewhat time consuming. The data administrators will need to find the oldest piece of tape hardware or software in the datacenter, then add up every tape associated purchase made since that hardware or software was first purchased:

- Physical tapes
- Tape drives
- Tape libraries
- Tape servers
- Hardware maintenance
- Backup software
- Any other tape associated costs

Now take that number, and divide it by the total amount of tape data moved over that time period.

In this case, we will save you the long-form calculations, and get to the bottom line: our customer research shows that costs typically range from $0.01-$0.03 per GB moved, or **$10.24-$30.72 per terabyte moved**.

Tape backup infrastructure also has **power and cooling costs**. Typically we see a range of energy requirements of 0.25-0.5 kWh per terabyte moved. Assuming the same $0.12 per kWh cost of electricity, we can estimate the unit cost at **$0.03-$0.06 per terabyte moved**.

Since that tape backup infrastructure sits in the datacenter along with the storage, there are **datacenter costs** as well. A rack unit is going to cost the same if it's occupied by a flash array or a tape library. This can get a little bit tricky, as the rack space required for tape will vary not only by the type of hardware, but also by the organization's recovery-time objectives. Some companies keep only a small percentage of their tapes in the tape library at any time in order to cut down on the data-center space required. The trade-off is that the time to recover data may be much longer, since the backup administrator may have to go digging through the closet to find the tape they need. Other companies might keep all their tapes in the library all the time, so they can be accessed as quickly as possible. The trade-off in their case is that this approach significantly increases their datacenter space consumption.

By weighing the difference in the two approaches, and by making some reasonable assumptions about each, we come up with a calculated range of **$0.48-$1.00 per terabyte moved**.

As a backup solution, tape is notoriously labor-intensive. Many companies hire an entire team of administrators just to manage the constant barrage of tape-related activities– swapping tapes in and out of drives, fielding restore requests, fixing tape streams that constantly seem to fail, rebooting hardware, updating software, and coordinating offsite tape rotation, among other tasks – making **personnel costs** one of the larger budget line items in a tape environment.

Most of these processes can't be easily automated, so in addition to being labor-intensive, tape-backup services are also notoriously error-prone, and one lost or damaged tape can mean the loss of an entire backup set. The infrastructure and labor costs, together with the risk and exposure of human error and media failures, are the primary reason why tape backup fails at scale, pushing many companies to look for other solutions. Assuming the fully burdened cost for an admin is still $120,000, your tape costs will be an additional **$0.75-$1.50 per terabyte moved**.

Adding it up, the total costs come to **$11.50-$33.28 per terabyte moved**. This seems extremely cheap, but the key distinction here is that this is per terabyte moved, which as we mentioned earlier is not the same as terabytes on primary storage. This means that the typical range for the cost of tape backup is **$138-$1,731 per terabyte per year!** If you are willing to pay extra to compress data as it is moved to tape, you may be able to reduce the total spend. However, you are still going to end up allocating a considerable amount of your overall IT budget just to tape costs. This is staggeringly high when compared to the unit cost of a tape drive, but demonstrates just how quickly tape costs can accumulate.

For the example company whose primary storage ecosystem we looked at above, their tape costs would escalate rapidly:

Year	1	2	3	4	5
TB stored on primary tier	1,229	1,475	1,769	2,123	2,548
Cost of tape hardware + software	$327,156	$392,587	$471,104	$565,325	$678,390
Cost of tape power & cooling	$831	$997	$1,196	$1,435	$1,722
Cost of tape rack space	$15,335	$18,403	$22,083	$26,500	$31,800
Cost of tape FTE	$23,962	$28,754	$34,505	$41,406	$49,687
Cost of tape backup	$367,283	$440,740	$528,888	$634,666	$761,599

Now let's look at replication

As unstructured data has proliferated over the last decade, many organizations have outgrown their tape backup solution's ability to protect their data. Those companies have implemented disk-to-disk (D2D) replication as a backup strategy. D2D solutions require vendor-specific replication software to deliver the performance needed to keep up with the rapid generation and churn of unstructured data at scale. Unfortunately, this same replication software requires backup workflows that replicate data from the primary NAS system to a secondary storage platform from the same vendor. In most cases, this means that the secondary storage system is architecturally similar to the primary NAS device, only slower and denser.

For replication purposes, a secondary storage tier typically doesn't require the same level of performance as the corresponding primary tier. It will need to be able to handle the amount of data coming in, and it should be capable of handling typical restore workflows, but those are usually less demanding than the production performance that end users need from primary storage. This means most organizations can purchase more cost-effective **hardware and software** for their secondary targets.

However, since backup datasets must maintain version histories of changed files, and because backup copies are often retained for months or years, secondary storage can often require far more capacity than primary NAS. For example, let's look at an organization that has one petabyte of primary data, with a yearly growth rate of 20%, a monthly change rate of 5%, and a three-year retention policy. After three years, they will have 1.4 petabytes of primary data, but they will need 3.1 petabytes of secondary capacity to hold all of the replicated data.

Many organizations report that they budget an amount equivalent to 80%-300% of their primary hardware and software budget for secondary hardware and software. Given our calculated range from our estimate of primary storage costs, this would translate into an estimated cost of **$140-$1,500 per terabyte per year** for secondary hardware and software.

Secondary tiers are effectively the same as primary tiers from a **power and cooling** standpoint. We can simply reuse our **$4-$6 per terabyte per year** estimate.

While typically less than the administrative overhead associated with tape backup, the amount of employee hours spent managing replication workloads is roughly equivalent to what's spent on primary storage. Assuming approximately $375 per rack unit, this gives us an estimated backup FTE cost of **$2-$4 per terabyte per year**.

That said, secondary storage targets are typically denser than primary NAS systems, and thus have lower **datacenter costs**. A typical 4U secondary system will have somewhere between 480 terabytes and 960 terabytes of capacity, giving us a datacenter cost of **$1-$2 per terabyte per year**.

Putting this together, we get an expected range for the total cost of replication of **$147-$1,512 per terabyte per year**.

Year	1	2	3	4	5
TB stored on primary tier	1,229	1,475	1,769	2,123	2,548
Cost of replication hardware + software	$322,560	$387,072	$464,486	$557,384	$668,860
Cost of replication power & cooling	$4,719	$5,662	$6,795	$8,154	$9,784
Cost of replication rack space	$1,475	$1,769	$2,123	$2,548	$3,058
Cost of replication FTE	$2,304	$2,765	$3,318	$3,981	$4,778
Cost of replication backup	$331,057	$397,269	$476,722	$572,067	$686,480

You can see why IT and business executives complain about how much they spend just on backup!

The total cost of data on primary

After all this, we can finally assess the true cost of storing unstructured data. By adding up all the hardware, software, energy, facilities, and staffing costs for primary and backup capacity, we see that a single terabyte of data, stored on a primary tier and backed up via tape or D2D replication, can cost an organization anywhere from **$322-$2,270 per year!**

Obviously that is a wide range, which is expected because different organizations have vastly different requirements for primary performance and backup strategy. However, even the lowest end of that range – $322 per terabyte per year – is likely much higher than what most companies believe they are paying.

The alternatives to keeping data on primary storage

When evaluating whether any infrequently-used dataset should remain on primary storage, we should weigh the exceptionally high cost of keeping data on primary storage against the other options available.

Every company needs to evaluate the cost of the storage alternatives in order to compare it against the current total storage costs. Assuming IT has the tools – such as Igneous DataDiscover – to determine which data is still capable of returning value to the organization, a decision must be made about what to do with the company's cold data – whether to archive it, or delete it.

While data deletion is simple and free, it risks removing files that could be needed in the future. If an old project is reopened, the cost to regenerate that data – if it can even be done – could be exorbitant. Additionally, in some industries, regulatory policies can prohibit deletion of data for many years, and regulator agencies can levy financial penalties for noncompliance. For organizations in those industries, deletion may not even be an option until years after the data is created. That time translates into a lot of money spent on mostly unused data.

Archiving data, however, offers some major advantages relative to primary storage, since hosting costs can be minimized while still maintaining access to data. Those same unused datasets that are costly to keep on a primary tier – but can't be deleted entirely – are comparatively cheap to store on an archive tier. From the data owner's perspective, the only sacrifice required is the turnaround time to move the dataset back to primary storage in the unlikely event it is ever needed.

Assuming archive of cold data is preferred to deleting it, there are two choices: on-premises storage or public cloud storage. Igneous DataProtect customers can choose between archiving their data to dense, cost-effective storage in their datacenter, or using their Igneous instance to move it online – including any tier of AWS, Azure Blob Storage, or Google Cloud Platform (typically, customers select the coldest tiers available from each of these providers for their archive storage). On-premises archive offers the advantage of nearly immediate recoverability of archive data, while the coldest tiers of cloud – such as AWS Glacier Deep Archive – are the cheapest options available.

Let's analyze the cost factors of archive storage to get a complete picture of the options.

Archiving on premises

Storing cold data on premises requires a lot of the same components of primary storage or replication, albeit at a much cheaper price point.

Hardware for Igneous on-premises storage involves two components: databoxes and application service routers (ASRs). Our typical databox provides 426 terabytes of usable capacity. For archive purposes, we'd usually recommend a 1:4 deployment, with one ASR managing up to four databoxes. All of the hardware is sold at-cost through distributors, without any markup by Igneous. A databox typically costs $32,320, while an ASR will cost $17,848. Assuming a 5-year lifespan for the hardware, this gives us a total hardware cost of **$17.26 per terabyte per year**.

One thing to note here is that this cost assumes full capacity utilization of a databox. If an organization has 500 terabytes of data, they would need two databoxes and one ASR, totalling 852 terabytes of capacity, and a hardware cost of $33 per terabyte per year. The cost per terabyte will always be higher because of the difference in capacity and consumption. However, using the cost per terabyte for capacity rather than consumption is simpler for this exercise.

The energy required to **power and cool** a full 1:4 arrangement with 1,704 terabytes of capacity is 67,027 kWh per year. At our defined electricity rate of $0.12 per kWh, this turns into a cost of **$4.72 per terabyte per year**.

A databox requires four rack units in a datacenter, while an ASR requires two. This means a 1:4 arrangement will occupy 18 rack units while delivering 1,704 terabytes of capacity. If a rack unit costs $240 per year, this calculates to **$2.54 per terabyte per year**.

Every Igneous product is delivered as-a-Service. This means that **personnel costs** are freed up within the organization to work on other IT projects that can deliver more value to the company. Routine, low-value tasks – such as managing hardware, updating software, monitoring tasks, and triaging issues – are all handled by Igneous. Additionally, the simple user interface, search-to-restore capabilities, and highly scalable, policy-driven workflows further reduce the amount of time required from IT admins. Our customers report that the employee hours typically spent on data management drops by 90% after subscribing to Igneous. Using the same cost estimates as before, this gives us an estimated personnel cost of **$0.40 per terabyte per year**.

Putting these together gives us a total cost of **$24.92 per terabyte per year** to store archive data on premises.

Year	1	2	3	4	5
Total data on archive	341	444	567	714	891
Cost of archive storage	$60,203	$32,321	$0	$0	$32,321
Cost of archive power & cooling	$3,052	$4,715	$4,715	$4,715	$6,379
Cost of archive rack space	$1,440	$2,400	$2,400	$2,400	$3,360
Cost of archive FTE	$135	$176	$224	$283	$353
Total cost of archive	$64,829	$39,612	$7,340	$7,398	$42,413

Archiving to public cloud storage

Cloud costs are notoriously hard to decipher. Costs vary by vendor, and every data transaction in the cloud seems to include a metered cost. Publicly-available pricing information show costs for storage, put operations, get operations, retrievals, and transfers, along with a laundry list of other functions without making it clear what you get charged for and when.

Fortunately, we've done the work and figured out how to make cloud pricing simple to understand. Additionally, we've engineered our product to move data to and from the cloud using processes that minimize costs for our customers.

Storage fees are actually pretty simple to enumerate when you are archiving to public cloud using Igneous DataProtect. Most cloud architectures require that data be written to a hot storage tier, then committed to the desired long-term tier afterward. This approach adds transaction fees to move data to archive storage. In contrast, Igneous writes directly to the desired tier, eliminating the extra transactional costs. Our method also reduces ingress fees to a single transaction per blob (more on blobs in a moment).

Writing directly to the desired storage tier makes calculating annual **storage costs** very straightforward: it's just the total size of the compressed data multiplied by the annual storage fee for the tier.

To keep our comparison easy, let's focus on the cost of storing a single terabyte in the cloud for one year. The following chart shows how much you can expect to pay in storage fees per terabyte per year for every cloud provider and tier if you've used Igneous to upload your data:

Tier*	AWS	Azure	GCP
Hot (S3, Hot, Regional)**	$258.05	$208.90	$319.49
Warm (IA, Cool, Nearline)	$153.60	$122.88	$122.88
Cold (Glacier, Archive, Coldline)	$49.15	$12.17	$86.02
Coldest (Deep Archive, Archive, Ice Cold)	$12.17	$12.17	$14.75

*All prices US West region

**Hot tiers on AWS and Azure have a sliding scale whereby the first terabyte is ~10% more expensive than the 500th terabyte. This is the cost after storing 500+ terabytes.

For comparative purposes, let's assume that for archive storage, organizations will choose the cheapest long-term storage solution: AWS' S3 Glacier Deep Archive tier. The cost estimate for storage fees is **$12.17 per terabyte per year**.

Ingress fees are one-time charges associated with placing data in the cloud, and are assessed for every individual put operation. For companies that have millions or billions of individual objects, these transactions have been cost prohibitive, with cloud vendors typically charging between $0.005-$0.05 per 1,000 put operations. A company looking to move one billion files to a public-cloud archive tier would be charged over $50K just to upload their data!

Ingress fees are incurred separately from storage fees, so the $50K number would be assessed on top of the amount of capacity consumed. It wouldn't matter if this billion-file dataset was 100 terabytes in size, or 100 petabytes. For many organizations, even large enterprises, this hurdle can be a show-stopper even before they encounter the challenging logistics of moving that much data to public cloud.

Igneous reduces these costs significantly by using our efficient data-movement engine to bundle large numbers of files into blobs and then upload each blob to the cloud as a single object, with a single put transaction. In this manner, we incur only one put charge for that set of files. With typical blob size being 100 megabytes, the following chart shows how much you can expect to pay per terabyte in ingress fees for every cloud provider and tier using Igneous DataProtect:

Tier	AWS	Azure	GCP
Hot (S3, Hot, Regional)	$0.05	$0.05	$0.05
Warm (IA, Cool, Nearline)	$0.10	$0.10	$0.10
Cold (Glacier, Archive, Coldline	$0.52	$0.10	$0.10
Coldest (Deep Archive, Archive, Ice Cold)	$0.52	$0.10	$0.10

Again, these are one-time charges, not annual costs. In order to get a fair comparison, these costs should be converted into an annual cost per terabyte. Although data could be stored in the cloud indefinitely – meaning the amortized annual cost would drop to $0 – that's not realistic. Let's use the same 5-year life cycle we used for the on-premises hardware. In practice, this means that after 5 years in archive, data can be expired and deleted (which is a reasonable assumption for comparative purposes).

Continuing to use the AWS Deep Glacier example allows us to get to our cost estimate for ingress fees: **$0.10 per terabyte per year**.

Egress fees are paid when data is retrieved from the cloud and returned to on-premises storage. These costs are a bit more complicated, as every get operation to retrieve a blob will incur a one-time charge, along with separate charges for retrieval and transfer, based on the total amount of data downloaded from the cloud. This is further complicated by having to predict how much data will need to be retrieved per year from archives, which will vary significantly by company, and may also vary from one year to the next within the same company.

The following table shows the amount charged to restore a single terabyte from each storage tier on each major public cloud provider using Igneous DataProtect (note that these are one-time charges based on restoring a full terabyte):

Tier	AWS	Azure	GCP
Hot (S3, Hot, Regional)	$51.92	$51.20	$81.97
Warm (IA, Cool, Nearline)	$61.45	$61.45	$92.26
Cold (Glacier, Archive, Coldline	$61.96	$76.92	$133.22
Coldest (Deep Archive, Archive, Ice Cold)	$61.96	$76.92	$133.22

Again, this cost is estimated per terabyte, and most organizations will very rarely need to recover every byte they put into archive. In any given year, our research indicates that a typical company will recover between 0%-5% of their total archive data, with the median recovery rate being about 3%. Under this assumption, for every terabyte archived to the public cloud, an organization can reasonably plan to recover about 0.03 terabytes per year. Again, assuming the data is stored in AWS Glacier Deep Archive, this gets us to our egress cost estimation: **$1.86 per terabyte per year**. This number, thanks to Igneous DataProtect's architecture, is likely to be significantly lower than what most IT departments would expect, given widespread concerns about the cost of recovering data.

Because Igneous services are all delivered as-a-Service, **personnel costs** will remain the same for either on-premises or cloud archive. Regardless of where your data is stored, we will be monitoring system health, triaging issues, and maintaining the service. This means we can re-use the **$0.40 per terabyte per year** estimate that was previously calculated.

Putting all of these costs together gives us an expected cost of **$14.53 per terabyte per year**. It's worth pointing out that the cost of storing data in the cloud is approximately half the cost of storing data in even the cheapest on-premises storage solution.

Year	1	2	3	4	5
Total data on archive	341	444	567	714	891
Ingress Fees	$179	$54	$64	$77	$93
Egress Fees	$635	$825	$1,053	$1,327	$1,656
Cost of archive storage	$4,152	$5,398	$6,893	$8,687	$10,839
Cost of archive FTE	$135	$176	$224	$283	$353
Total cost of archive	$5,101	$6,452	$8,235	$10,374	$12,941

Igneous subscription costs

Igneous DataProtect costs the same for our customers, regardless of whether they opt to store archive data onsite or in the public cloud. The prices for our services are straightforward.

Igneous DataProtect is priced on a per-terabyte-under-management basis. For archive customers, the list price is **$200 per terabyte per year**, with discounts becoming available as capacity under management grows.

All Igneous DataProtect customers can take advantage of:

- Simple search-to-restore across all archived data, regardless of location, that provides visibility into even the largest scale environments
- Source-agnostic platform that can archive data from any NFS or SMB source
- Intuitive interface to archive whole datasets with only a few clicks
- As-a-service delivery supports a comprehensive data-management strategy with nearly no operational overhead
- Latency-aware data movement to ensure archive operations create no impact on end users or production workflows

The Igneous DataProtect service allows our cloud customers to benefit from:

- Universally compatible platform that enables archiving to any public cloud vendor and tier
- Massive cost savings for ingress and egress of data to and from public cloud storage
- Simple process to setup and execute movement of data to public cloud
- Automated cloud capacity control with our state-of-the-art cloud expiration functionality

To be conservative, let's use the list price and assume no discounting.

The benefits of compression

A major benefit of archiving data, either on premises or in the cloud, is compression. Primary data is stored in native format so that end users and applications don't have to do extra work to use it. Archive data, on the other hand, can be compressed down to the smallest possible footprint to fully minimize space consumption and maximize savings, since there is no expectation of performance for data that is not in active use. This isn't new technology, but it can add up to significant savings.

Compression ratios vary from 1:1 (for incompressible data) up to 2.5:1 or higher for more compressible workflows. To get to a real dollar cost, let's assume a compression ratio of 2:1. This further reduces the cost of archive storage by 50%, since a terabyte of cold data on primary storage will only require 0.5 terabytes in its archived state.

The total costs of archive

Adding all of the costs above and accounting for compression, the total costs for archiving cold data to the cloud comes out to **$114.53 per terabyte per year**. Archiving on premises is slightly more expensive, at **$124.92 per terabyte per year**.

Walking through an example

After laying all of these numbers out, it becomes clear why an effective and ongoing archive strategy is a huge advantage for organizations with lots of unstructured data. Every unused terabyte that is discovered and archived using Igneous DataProtect could save **$195-$2,028 per terabyte per year!** This represents a savings **of 63%-94%** compared to the total cost of storing data on a primary storage tier and backing it up using traditional tape or D2D replication workflows.

Igneous DataDiscover customers typically find that 40%-60% of their data currently on primary storage hasn't been touched in over a year. For a customer with a petabyte of unstructured data, this could mean that over 600 terabytes could be safely archived, representing a savings opportunity of **$117K-$1.2M per year**, or **$585K-$6.1M over five years**. That is a massive amount of money to leave on the table.

Additionally, while unstructured data typically grows by 10%-40% per year, the amount of data in use at any one time doesn't increase at the same rate. If a customer with a one-petabyte data footprint is growing by 20% per year, they would expect to add about 205 terabytes in the next year, bringing their total unstructured data on primary to 1.2 petabytes. However, the amount of data in use at any time is highly unlikely to grow by 205 terabytes. As data is generated, it replaces old data that is no longer in use.

Typically, the amount of data in use grows at approximately the same rate at which overall data footprints grow. This means that in any given year, the amount of data that becomes cold is approximately equivalent to the amount of data generated in the previous year. If an organization with one petabyte of unstructured data grows their data by roughly 20% per year, they would expect to add 200 terabytes in the first year, and 250 terabytes in the second year. However, in the second year, they can also expect to have approximately 200 terabytes go cold, meaning their active data only grows by 50 terabytes.

If this organization continues to archive datasets that age out of active use, they can shrink their primary capacity consumption growth from 20% to somewhere closer to 5%, meaning an additional savings of **$39K-$406K in the first year**. Their annual savings would continue to increase by at least this much over five years as more cold datasets continue to be archived.

Example company savings

So what is the total cost of cold data? Let's take a comprehensive look at the example organization we've been discussing, and make the following assumptions about their environment and operations:

- Primary data: 1 petabyte
- Growth rate: 20% per year
- Primary hardware and software costs: $175 per terabyte per year
- Backup strategy: tape
- Tape costs: $0.01 per gigabyte moved
- Tape level-0s per year: 52 (1 per week)
- Archive strategy: Cloud
- Archive recovery rate: 3% per year
- Cloud vendor: AWS
- Cloud tier: Deep Glacier
- Compression ratio: 2:1

Let's also assume they have done everything they can to minimize costs in their existing infrastructure, such that current primary and backup costs are on the lowest end of the normal range.

Primary capacity

Without archiving any of their cold data, this company would start with 1,024 terabytes on primary storage, and would need at least 2,548 terabytes of capacity to handle their growth by the end of Year 5. Igneous DataDiscover shows this company has an average amount of cold data, with 500 terabytes having gone unused over the last year. With an archive strategy that moves those 500 terabytes to cold storage, they will need only 524 terabytes of primary NAS capacity to start, and will need only 778 terabytes of primary storage by the end of their fifth year.

This means that this company could *delay purchasing additional primary capacity for at least five years*. Budget currently allocated to expand primary capacity multiple times and maintain aging infrastructure could instead be spent upgrading existing primary capacity, with plenty left over for other projects.

The total cost of the primary environment - including hardware, software, power and cooling, datacenter costs, and FTE costs - would start at $226K in Year 1 and grow to $469K by the end of Year 5 if an archive strategy isn't implemented. With Igneous DataDiscover identifying datasets for archive, and Igneous DataProtect archiving these datasets to AWS, the costs would drop to $100K in Year 1 and $141K in Year 5.

The total cumulative cost savings for the primary tier alone would be $1.1M over 5 years.

Backup costs

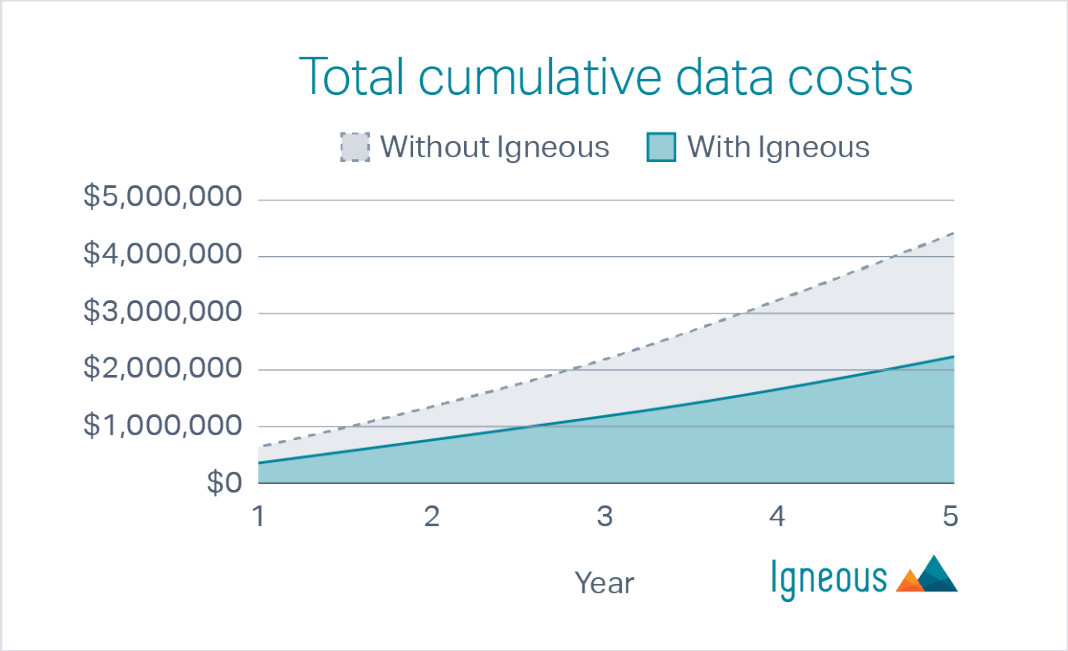
This same company would normally expect to spend $367K on backup hardware, software, power and cooling, datacenter costs, and FTE costs in the first year. That would grow to $762K by the end of Year 5. With Igneous DataProtect and Igneous DataDiscover archiving their data, those numbers plummet to $167K in the first year, and only $233K in Year 5. How many IT administrators have ever managed a budget that needed nearly 40% less backup capacity than five years ago even while their data is growing at 20% per year? The cumulative savings from backup would be $1.77M over the next five years.

Archive costs

Using Igneous, this company would instead spend only $5K with AWS in the first year, growing to $12.8K in Year 5. The cost of the Igneous subscription would total $67K in Year 1, and top out at $177K in Year 5. These new costs are only a small fraction of the overall savings from reduced primary storage consumption and backup.

Cumulative costs

Our calculations show that this company would save **$2.22M over the next five years** by using Igneous to find and archive their cold data. The total cost of doing nothing and continuing with business as usual would be $4.4M over the next five years. The total cost of unstructured data infrastructure including primary, backup, archive, and Igneous would be $2.19M over that same timeframe. This translates to a total cost reduction of over **50%**.



Conclusion

IT administrators can pair this information about the true cost of cold data with the real-time data insights provided by Igneous DataDiscover, which can quickly and easily identify datasets that haven't been touched. That is an extremely compelling story to take to data owners: it identifies all the datasets that are no longer in use, then shows them why it's critical to archive those datasets. Combine this with the ability of Igneous DataProtect to provide direct read-only access to archived data, or to restore data from on premises or public cloud to the primary tier quickly, and any rational argument for keeping cold data in place no longer adds up.

Costs vary significantly from one industry to the next, and even within industries every organization is different. To account for this, we have provided ranges from the lowest to the highest costs we have seen in practice in addition to providing you with estimates for averages in a typical enterprise.

Contact Igneous

Igneous has built a customizable, sharable calculator based on the information contained in this whitepaper. We encourage you to use your numbers in this calculator to get an estimate of your true cost of cold data. To get access to the calculator and view a customized estimation on your costs of cold data based on your organization's specific data profile, contact us at info@igneous.io or 844-IGNEOUS.

2401 Fourth Ave, Suite 200, Seattle, WA 98121, USA / 1-844-IGNEOUS / www.igneous.io

Igneous