WINTER CORPORATION

Modernize Hadoop Data Stores with DataOps

WINTERCORP REPORT

BY RICHARD WINTER

SPONSORED
RESEARCH
PROGRAM

EXPERTS IN ANALYTIC
DATA MANAGEMENT AT SCALE

WINTER CORPORATION

Modernize Hadoop Data Stores with DataOps

RICHARD WINTER

WinterCorp
www.wintercorp.com
42 DERBY LANE, TYNGSBORO, MA
617-695-1800

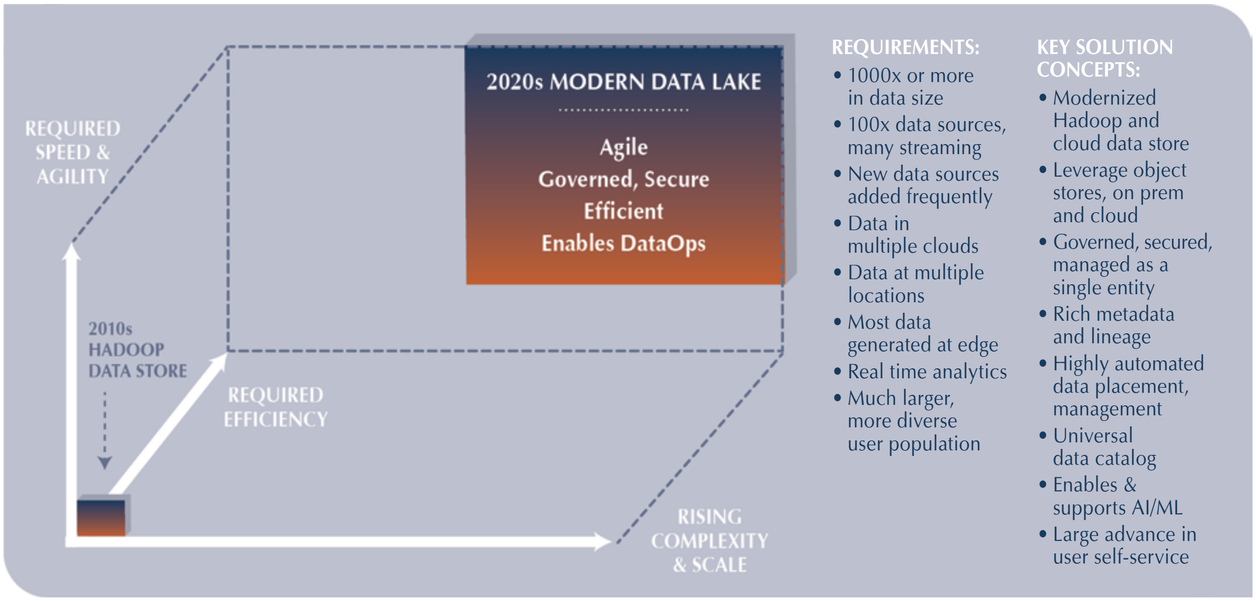
EXPERTS IN ANALYTIC DATA MANAGEMENT AT SCALE

MOST EXISTING BIG DATA STORES, whether on Hadoop or in cloud storage, and whether or not they are called "data lakes," are disorganized dumps of files in which it is difficult to find anything; there are many copies of the same data; data quality and lineage are unknown; little or no data is actually curated; little data is shared or reusable; and, because of inadequate security, the entire collection of data poses one massive security risk. This is the proverbial data swamp and every year there are more of them. Worse, most of those that exist continue to grow because customers have found no better way to deal with their always growing volumes of "big data."

As if these problems were not enough, many executives are also finding they don't care much for the annual Hadoop upgrade. If your data lake is on premises, in one or more Hadoop clusters, you probably have to add more nodes every year. Not because you need more computing power, but because you need more data storage. And, now that the cluster has grown large, that data storage doesn't look as inexpensive as it once did, in part because you are paying to store three copies of each file. In order to add storage capacity, you must add entire servers, paying for processors and memory in addition to each unit of storage.

And, while some vendors tout "cloud" as the simple solution, not everyone finds it is economical to store all the "big data" in the cloud. Object storage in the cloud is low cost only if you don't actually access the data. If you use the data intensively, the access charges mount rapidly. Moreover, storing large amounts of data in the cloud often give rise to bottlenecks at access time.

The time has come to rethink the Hadoop data store, to address these problems and the coming data issues of the 2020s. This report discusses today's challenges for the Hadoop data store; discusses the principles of a new approach based on DataOps, a set of principles and methods aimed at increased agility; and, makes recommendations. ●



REQUIREMENTS:

- 1000x or more in data size
- 100x data sources, many streaming
- New data sources added frequently
- Data in multiple clouds
- Data at multiple locations
- Most data generated at edge
- Real time analytics
- Much larger, more diverse user population

KEY SOLUTION CONCEPTS:

- Modernized Hadoop and cloud data store
- Leverage object stores, on prem and cloud
- Governed, secured, managed as a single entity
- Rich metadata and lineage
- Highly automated data placement, management
- Universal data catalog
- Enables & supports AI/ML
- Large advance in user self-service

Table of Contents

1 Introduction

A company that wants to compete in the 2020s is likely to be using data and analytics as a core component of its strategy and operations, leveraging a wide range of data via data science, machine learning and artificial intelligence. Such a company needs data that is comprehensive, appropriately curated, secure, accessible and economically managed. To compete in the 2020s it will also need agility with data and analytics to accelerate data-related projects.

The Hadoop data store plays a key role in the analytic data strategy of a modern enterprise. The Hadoop data store is the only place in which the total data assets of an enterprise can be managed. In contrast, the data warehouse continues as a strategic component but is usually focused on only the most intensively used data at the core of business operations. The Hadoop data store is meant to encompass ***all*** of the data of continuing value in the enterprise — typically between 10 and 100 times as much data as in the data warehouse, and a far more diverse collection of information.

Because the Hadoop data store is more comprehensive in both its contents and its tooling, Hadoop data stores support empirical insights from data that serve to complement an enterprise data warehouse (EDW). In addition, the Hadoop data store typically has a lower system cost per unit of compute or storage than the data warehouse, providing an opportunity to advantageously offload from the data warehouse. Processing offloads are usually compute intensive tasks, including ETL and deep analytics. Data offloads are usually data sets with lower value density than the core enterprise data that is often retained in the EDW: data such as videos, audio, scans and other data that is valuable to analyze but whose value is not enhanced by placing it within a relational database structure.

Many of the key business strategies of the coming decade — whether they exploit sensor data, image data, speech data, web data or more traditional data sources — will rely on the existence of a well-managed Hadoop data store in which massive volume is accommodated with a cost-effective infrastructure that puts both cloud and on premises data storage to its best use.

> **In the coming decade,** most existing Hadoop data stores are going to have to be transformed into an entity that enables more effective data governance, self-service use by a much larger audience, more cost effective operation and much higher operational agility. Via a process of modernization, a Hadoop data store can be transformed into a modern data lake for the 2020s. ●

This WinterCorp Report is about that process of Hadoop modernization and its end result: the modern data lake.

2 Hadoop Data Store Challenges

Thousands of so-called data lakes have been implemented over the last ten years but few have fulfilled their promise. According to Gartner, only 15% of data lakes have even reached production status.[1] Worse, most continue to expand, costing their owners more each year, while delivering value only in a narrow sense and producing little or no return on investment (ROI).

Many executives started out with the mistaken idea that simply installing a Hadoop cluster and making it available to users would give them a usable data store. They discovered that every major increase in Hadoop storage capacity requires the addition of compute capacity, resulting in large, rapidly growing system cost. Then, when cloud object storage became popular, many or all of the files in these dysfunctional data lakes were simply copied to the cloud.

[1] "How Can I Help?", Nick Heudecker, Gartner Blog Network, https://blogs.gartner.com/nick-heudecker/how-can-i-help-breaking-down-the-dallas-data-analytics-conversations/

Unfortunately, a data swamp remains a data swamp when you move it from Hadoop to the cloud. IDC estimates that only 1% of data stored by enterprises is actually used.[2] The under-utilization is not because the data lacks value — it is because the data is too hard to find and too hard to use. Referred to as "dark data," much of this remains opaque to business users who have an interest in applying it to derive business value. When an un-architected, un-managed data store is moved to the cloud, one still has a very inefficient, error-prone operation. Costs are often higher after the move because **frequently accessed data costs more, not less, in the cloud**.

The root problem is that people underestimate the impact of having the data store grow to a scale that is unprecedented in the organization. People are not used to the vastness of this ocean of data: informal methods of curation and management no longer work when there are millions or billions of files.

Even in so-called "successful" data lakes, 50% of project effort is devoted to data preparation: finding the data; cleansing it; formatting it; and, integrating it with other data needed for the project.[3]

What is needed instead is an approach in which the essential practices happen automatically because they are built into the processes for using the data lake.

3 Data in the Modern Enterprise

The previous section summarizes what we hear when we talk with Hadoop data store owners about their experience in recent years. While these are important, there are even larger challenges ahead.

Since the data lake concept emerged about ten years ago, the world has changed significantly with respect to "big data." The key changes affecting the data lake are: speed of business; new and much more numerous data sources; much more intensive use of certain data sources, such as sensor data; remarkably more intensive use of data science, machine learning and artificial intelligence; self-service for business analysts; and, large increases in scale. Each of these changes serves to intensify the challenges faced in building and maintaining a data lake, thus creating even more pressure to find a more streamlined and economical approach to the big data.

Overall, what stakeholders need with respect to their data is:

- **An acceleration of project success: more rapid time to value;**
- **Control of their data management costs; and,**
- **Business-appropriate control of risks, including data security and privacy risks.** ●

These and other key data-related requirements are discussed below.

3.1 SPEED OF BUSINESS

The modern enterprise must compete in a much faster moving world than we had just ten years ago when Hadoop first emerged.

The data store must quickly accommodate change; it must facilitate rapid prototyping and experimentation; it must support rapid implementation at scale of new data-based business solutions; and, it must incorporate continuous rapid changes to large volumes of data about customers, their behavior, their opinions about companies and products — and their interests and desires.

This need for speed and agility requires advances in architecture; new product capabilities; and, new and more varied business processes around the data lake and its governance. ●

[2] "IDC: The Digtal Universe in 2020", https://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm

[3] "Consuming Big Data: Most-Time Consuming, Least Enjoyable Data Science Task, Survey Says", Gil Press, Forbes, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#ea9b9516f637

3.2 NEW DATA SOURCES

Just as the speed of business has advanced rapidly, the number of new data sources relevant to a business has increased at an even greater pace — and that pace is continuing, if not accelerating.

Whereas the early Hadoop cluster typically drew data from a few dozen operational systems within the company, modern data lakes are now often receiving data flows from thousands or tens or thousands of sources. As well as the tabular, highly structured data typical of relational databases, the newer sources are often semi-structured records (e.g., sensor data, XML, JSON) and so-called unstructured data (e.g., free-form text, video, images or audio). Examples of this include sensor data (IoT), social media data and log data from websites and computer systems.

> **Taken together,** there are many more sources of data relevant to corporate decision making than there were in the early stages of data lake operation. This is a trend likely to continue and accelerate in the coming decade. ●

3.3 DATA SCIENCE / ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

The advanced statistical methods of data science and the models associated with machine learning were not initially seen as relevant to every business.

But now most businesses employ data scientists and machine learning specialists or consume these services from outside providers. In either case, the data that is needed for their models is often in the Hadoop data store.

In general, the rising significance of machine learning, artificial intelligence and data science imposes more particular standards of curation on the data lake. Note that only a small percentage of machine learning models developed in recent years have actually been deployed into production.[4] **In the decade ahead it will be increasingly important for the data lake to enable successful development and deployment of machine learning.**

There are three critical factors here. **First, you need to find the needed data**, which requires a strong catalog; often tagging; and, a good search engine. **Second, machine learning often requires data that has higher quality and that has more complete lineage** than most data found in most Hadoop data stores today. **Third, successful machine learning often requires a greater quantity of relevant data, or more detailed data**, than is present in many Hadoop data stores today. What is present in the way of relevant data is often limited by either the economics of storage or by the cost of sourcing or ingestion.

Besides governance, end-to-end data lineage provides scientists and analysts with greater visibility into data sources, transformation history and rules that may have been applied to data. This is to determine the source and validity of the findings and recommendations of machine learning engines. Whether expected or anomalous, the outcomes of machine reasoning often must be traced, explained or analyzed.

[4] Towards Data Science Blog, "Why is Machine Learning Deployment Hard?", Alexandre Gonfalonieri, https://towardsdatascience.com/why-is-machine-learning-deployment-hard-443af67493cd

Purpose and Methodology for this Report

*This **WinterCorp White Paper** describes the modern Hadoop data store and its significance to customers for data management. In developing this report, WinterCorp drew on its own independent research and experience, interviewed Hitachi Vantara employees, attended company events and analyzed product documentation. Hitachi Vantara was provided an opportunity to comment on the paper with respect to facts, in its capacity as the sponsor of this research. WinterCorp has final editorial control over the content of this publication and is solely responsible for any opinions expressed.*

WinterCorp

It is often because these three critical data factors have not been addressed in existing Hadoop data stores that the machine learning models do not perform well, fail to inspire confidence and then fail to be deployed. ●

3.4 SELF-SERVICE

In the modern enterprise, business analysts who work hands-on with data increasingly want the capability to bring data they choose — perhaps from an external source — into the data lake environment and then combine it with other data.

Traditional IT departments have not typically accommodated this need. Rather, they have assumed that only professionally trained IT specialists can bring new data into the enterprise environment. In doing so it is expected that the IT professionals will follow rigorous, often delay-prone processes to ensure security, data integrity, correct handling of data rights, availability of support resources and so on. While this is appropriate for the most widely used, base core data of the enterprise, such processes need not be applied ahead of the exploration of new data in advance of a single experimental use.

For some time, companies have needed a self-service capability for end users — business analysts, data analysts and data scientists — to experiment with combinations of their own data and production business data, in an environment in which they don't create security risks and don't disturb production operations. ●

3.5 EXTREME SCALE

One of the most dramatic changes in the Hadoop environment since its emergence around 2010 is the growth in scale.

The largest publicly disclosed Hadoop clusters in the world in 2010 contained a few hundred petabytes of data, where a petabyte is a thousand terabytes. Today, these same systems contain tens of *exabytes* of data, where an exabyte is a thousand petabytes. Thus, data volumes are a few hundred times larger than they were ten years ago. In the next ten years, this trend is expected to continue to accelerate, with data volumes likely to increase yet again by hundreds of times or more.

So that is the challenge of the 2020s: how to manage yet more massive, more varied and more rapidly growing collections of data, somehow affording the data storage and providing all the services needed to make a successful Hadoop data store. That is typically going to grow to be **1000 times larger than current data volumes.** ●

3.6 CLOUD

The adoption of cloud computing is probably the biggest single change in the information technology field in the last ten years. Cloud revenues are estimated to have topped $227 billion in 2019, up over 15% from the prior year.[5]

Approximately 10% of all data centers were closed in 2017[6], even as investment in information technology continued to grow at a rapid pace. There is no question that many Hadoop data stores have been created in, or moved to, the cloud.

But the data management requirement is more than simply "run in the cloud."

[5] Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17% in 2020, https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020

[6] Mark Hurd, OracleWorld Keynote Address, 2018.

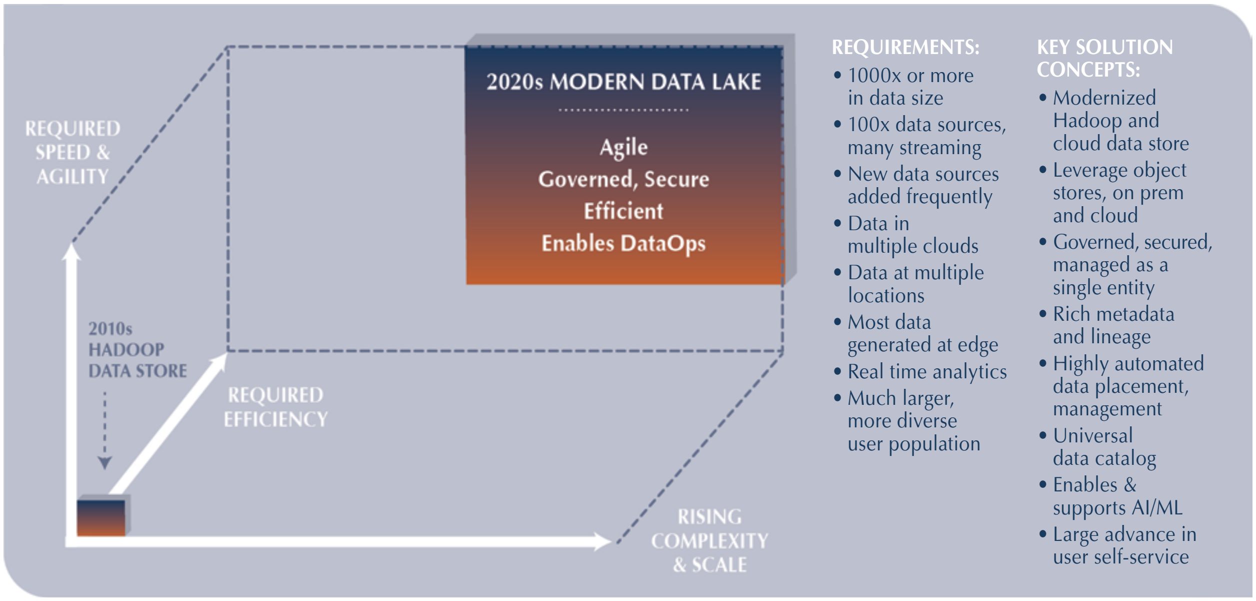
The modern data store needs to provide for cloud, multi-cloud, hybrid and on prem architectures. Only if you can advantageously offer all the options can you meet the needs of most enterprises for the decade ahead. And, in addition to deployment flexibility — meaning that the owner can readily choose between on prem and cloud deployment of analytics — it is going to be necessary to support the storage of data in multiple physical locations. Thus, the typical next generation data lake may well have data stored in multiple "on premises" data stores — each at a different location — in multiple regions of a public cloud — and in multiple clouds.

There are several forces mitigating against the storage of all of a company's data in a single location. **First, generation of extremely large volumes in the "edge"** (which itself is distributed around the world); **second, data sovereignty**: the idea that data generated in a given country may be required to remain in that country; **third, data gravity**: the idea that data which flows rapidly and intensively among a cluster of systems may need to be located near those system for either performance or cost reasons. Any of these forces may be outweighed, in certain circumstances, by other factors. But, in an enterprise that has customers, suppliers and business operations in multiple locations, it is unlikely that a single physical repository will successfully meet data lake requirements over the coming decade.

> **In fact, the modern data lake requirement** goes beyond simply providing for these deployment options to providing for effective data ingestion, governance, access, preparation and curation across the multiple environments, regardless of whether the data is on prem on in any of several clouds. The same principle applies to policies regarding data security, data quality, data access, data retention and data backup. ●

4 The Modern Data Lake Concept and DataOps

As we enter the 2020s, we need more than just a data store — a place to put data until we come back to it — instead we need an agile, governed, operationally efficient data entity. We need a modern **data lake** that fully addresses and supports **DataOps**, is an automated, process-oriented methodology, used by analytic and data teams, to improve the quality and reduce the cycle time of data analytics.



REQUIREMENTS:

- 1000x or more in data size
- 100x data sources, many streaming
- New data sources added frequently
- Data in multiple clouds
- Data at multiple locations
- Most data generated at edge
- Real time analytics
- Much larger, more diverse user population

KEY SOLUTION CONCEPTS:

- Modernized Hadoop and cloud data store
- Leverage object stores, on prem and cloud
- Governed, secured, managed as a single entity
- Rich metadata and lineage
- Highly automated data placement, management
- Universal data catalog
- Enables & supports AI/ML
- Large advance in user self-service

This support for DataOps must include: an automated, process-oriented set of practices used by analytics and data teams, to bring agility, timeliness, business value and reduced cycle times to analytic data projects. DataOps applies to the entire lifecycle of using data and recognizes the interconnectedness of developing solutions and delivering value through their production use. The goals of DataOps are similar to those of the widely successful practices of DevOps, an approach that has brought a new level of agility to the development of software applications.

To summarize what is needed in the modern data lake and to enable DataOps:

4.1 GOVERNANCE

A. Metadata. Metadata is data about data: when a file was last written or changed; what it contains; how the data is formatted; whether the data is encrypted; where it came from; and, so on.

Complete metadata must be captured and retained for every file that is ingested and for every data transformation or copy that occurs after ingestion, thus keeping track of datasets, data flows, policies, lineage and access control for the full lifecycle of the data object. This is the first, most essential, step toward a modern data lake. ●

There are two major ways to capture metadata: (a) get the metadata from the process as the data operation is performed (e.g., when a file is ingested, capture its source); (b) get the metadata at some later time, from crawlers that examine each object stored in the data lake, integrating this new metadata with any previously obtained metadata. In general, a combination of these techniques is needed, as there will often be a large, pre-existing collection of data prior to the imposition of an advanced metadata management regime.

Regardless of the approach, successful capture and retention of comprehensive metadata requires a systematic, focused approach. One requirement that may be new to many is this: in the modern data lake, the metadata grows massively. A petabyte scale data lake — which will be common in the coming years — will often have trillions of metadata records. The metadata is often changed and accessed more frequently than the data itself. So, performance and scalability are at least as important for metadata as they are for the data contents.

B. Universal Data Catalog. The metadata must be organized into a searchable, taggable, universal data catalog. Thus, the name and characteristics of every file, table or other data object stored in the data lake must be present in the searchable catalog — and the catalog must have advanced search capability on all data attributes.

If the data lake spans more than one platform — for example, if the data lake contains both data that is stored on premises and data that is in one or more clouds — then all of the data, regardless of deployment platform, must be represented in the catalog. The user must be freed from concern about where — in which data store or in which cloud — the needed data is stored. ●

The data catalog must be integrated with the access control and data security mechanisms so that users can access only data for which they have the required authorization.

C. Multiple Levels of Curation. There must be support for multiple levels of curation. Some uses may require raw data, exactly as received from the source, as is sometimes required in data science or auditing.

In other uses, such as business intelligence, the data will typically need to fully cleansed, transformed and integrated into a data model consistent with how the particular business function expects to see the data.

There are typically multiple different curation regimes in place for a data lake; the data lake should allow the customer to define and apply as many different curation processes and standards as may be needed. ●

D. Schema. When a schema, defining the structure of a record in the file, is present, it is part of the metadata and is retained permanently with the lineage. Any changes to the schema are similarly captured and retained, along with the date, time and circumstances of the change.

E. Security/Access Control. Data security and access control must be similarly automated on the basis of defined data attributes and policies. Such policies will determine where data can be placed (e.g., some sensitive data can be stored only on premises), whether it is encrypted, masked or anonymized; and, who can read it or change it.

4.2 AGILITY

A. Data Storage Cost. **Real agility in analytics is impossible without low cost, abundant storage.** This was one of the major drivers of the Hadoop architecture when it emerged about 10 years ago. Companies needed a way to store and manage data for analytics that was less costly than the data warehouse, which at that time often involved a large premium for data storage.

But, today, commercially supported Hadoop requires a software subscription for each node, which can range up to a list price of $10,000 per node. In addition, the standard Hadoop architecture couples compute and storage, so that one must purchase additional servers in order to get additional storage. Finally, each data file gets stored in triplicate, further increasing the overhead. While all of these costs may have been manageable when clusters were smaller, today they add up to a headwind discouraging new initiatives in analytics, especially when new data sources are involved. The rapidly increasing volume of data in many areas of business, due in part to the falling cost of sensors and other devices that generate digital data, are adding to pressure on storage cost; thus storage cost comes to hinder innovation.

Fundamentally, companies need a lower cost, readily expanded form of data storage. Industry has provided exactly this in the form of object storage. Object storage is available now in the cloud and on prem. The modern Hadoop environment will incorporate such object storage so as to economically accommodate the data and analytic needs of the 2020s. ●

B. Automated Data Lifecycle Management. Governance policies must be defined and automated. For example, there is likely to be a class of data for which a copy of the file is archived on receipt from the source, prior to any transformation or use. This must happen automatically. For some data, it may be important to encrypt the archived copy; this, too must be defined in the policy and automatically applied.

Data retention/deletion policies must be defined and automated. If, when data grows cold, it is to be moved to lower cost storage, this must be automated. If the data is to be deleted from the lake after a specified period, this must be in the policy and then be automatically performed. ●

C. Deployment Options. Contrary to a common perception, cloud storage is not always less expensive than storage on premise.

Cloud vendors charge not only for storing the data but also for accessing it and delivering it where needed. **Therefore, the real cost driver is the *total cost of storing, accessing and delivering the data to its users over its lifetime.*** ●

To give one example, an intensive pattern of access can make on premises storage cost far less than cloud storage.

Therefore, to optimize cost, the data lake owner must have the option of storing data in any of several clouds the company may use; and, any of several on premises data stores the company may use. Further, each of these data stores may be in a different location (though clouds are nominally without location, each cloud has "regions" and the choice of region may affect the cost of access), which can affect the total cost of using the data over its lifetime.

To effectively manage the costs of storing massive and rapidly growing collections of data, it is necessary for the customer to have the option of storing data on its lowest cost platform and location, considering all relevant costs. At the same time, data that is used intensively must be kept in a data store that provides the necessary performance.

Because there are so many data objects involved — billions or trillions in a sizable data lake — the data placement must also be policy-based and automated, where the inputs to the placement decision include data security, cost factors, sovereignty and service level requirements.

4.3 USER SELF-SERVICE AND ANALYTICS

The typical data lake has expanded well beyond the scale at which IT professionals can be responsible for ingesting all data, curating all data or implementing all the queries, reports and analytics required by users.

In most companies, it will be necessary for these functions to be handled by data analysts and business analysts in the business organization, except in particularly complex or sensitive situations.

User self-service requires that it be easy to do the ordinary, every-day functions: to ingest data; to do common transformations; to cleanse and profile data; and, to prepare most types of data for use; and, to perform queries and straightforward analyses. The involvement of IT professionals, using professional tools, may still be necessary for complex and demanding ETL processes. However, many other uses of data do not involve such complexities and people in the business functions can implement them, given sufficient automation, policy support and technical support.

The modern data lake must enable this data self-service paradigm. ●

5 Conclusions & Recommendations

Most Hadoop data stores, whether or not called data lakes, have failed to deliver business value. The data stored in these Hadoop clusters is almost entirely dark, unused and ungoverned. Prospective users can rarely find the data that they need; when they find data that may be relevant to their interest, they discover that its lineage and level of curation are unknown. As a result, it is either unusable or very costly to use. Many data lakes contain sensitive data that is unlabeled and uncontrolled, thus turning the data lake into a data breach waiting to happen. Even in the presence of all these problems, data continues to accumulate into these data swamps, often forcing their owners to continue to invest in upgrades.

The Hadoop data store of the 2020s must address and resolve these problems, as well as address the data and analytic challenges of the coming decade. The keys to a modern Hadoop data store for the coming decade include: automatic processes to capture and curate the necessary metadata and lineage; automated support for governance processes aligned with business goals and applicable regulations; enablement for DataOps; automatic placement of data across tiered storage options on premise and across multiple clouds; data catalog support; automation of data security and encryption; and, automated management of the lifecycle of data assets, including archiving and deletion as appropriate.

WinterCorp has created a Total Cost of Operation Model (TCO Model) that estimates the magnitude of the savings resulting from the use of an object store, for a given configuration, on the basis of options selected by the user of the model.

As an example of the savings that can be achieved on premise, the TCO model shows that a 25% expansion of a 10PB Hadoop cluster, including a disaster recovery cluster, will cost 75% less over five years with object storage, when compared to HDFS. **This is a savings of about three million dollars ($3M) in estimated cost.** Savings in space and power are similarly dramatic. The estimate includes hardware, software, labor and environmental cost factors. ●

A user of the model can adjust the inputs to reflect the situation of a particular customer or location.

On the basis of its independent research and consulting experience, WinterCorp recommends that companies investing in a modern Hadoop data store or data lake look closely at the concepts in Section 4 of this report and adopt them in their data store architecture for the 2020s.

Key among these concepts are adoption of:

- **DataOps, an agile set of practices for analytics**
- **Operational efficiencies, embodied in object storage; cloud; a range of cloud, hybrid and on prem deployment options; and automated data lifecycle management based on business driven policies**
- **A pro-active approach to governance encompassing a data catalog, rich metadata, lineage and multiple levels of curation**

Such a data strategy will enable a company to take the next crucial steps in its campaign to address the deluge of data and analytic requirements of the business environment of the coming decade. ●

WinterCorp is an independent consulting firm expert in the architecture and strategy of the modern analytic data ecosystem.

Since our founding in 1992, we have architected and engineered solutions to some of the toughest and most demanding analytic data challenges, worldwide.

We help customers define their data-related business interests; develop their data strategies and architectures; select their data platforms; and, engineer their solutions to optimize business value.

Our customers, with our help, create and implement cloud, multi-cloud and hybrid cloud architectures; they create the data foundation needed for data science, artificial intelligence and machine learning.

Our customers get business results with analytics in which their return is often ten or more times their investment.

When needed, we create and conduct benchmarks, proofs-of-concept, pilot programs and system engineering studies that help our clients manage profound technical risks, control costs and reach business goals.

We're expert with structured data, unstructured data, and semi-structured data — with the products, tools and technologies of data management for data analytics in all its major forms.

With our in-depth knowledge and experience, we deliver unmatched insight into the issues that impede scalability and into the technologies and practices that enable business success.

WinterCorp
www.wintercorp.com
42 DERBY LANE, TYNGSBORO, MA
617-695-1800