Technology Insights...

April 5, 2023

J.Gold Associates, LLC. Northborough, MA 01532 USA
www.jgoldassociates.com +1-508-393-5294
Research, Analysis, Strategy, Insight

Moving on from Outmoded Hadoop

A PUBLICATION OF J.GOLD ASSOCIATES

Hadoop was initially released in 2006 to process what was then an emerging problem – how to acquire, store and process very big data sets. Its inventors' primary application focus was data acquisition and processing across a distributed network of servers for search engine optimization. Hadoop divides its large data sets into smaller chunks to be run in parallel on multiple nodes forming a Hadoop cluster. Hadoop was built as an open source project using Java for its programming language and it enabled a democratized "Big Data" era with its distributed storage and processing capability.

Despite the Hadoop legacy solution being around for 17 years, albeit with some updates along the way, many Hadoop deployments still exist, although they are becoming less common. If you are one of the companies still using Hadoop, you should be asking yourself if that's the right approach given the changes in modern data processing and hybrid computing requirements. We believe that the cons greatly outweigh the pros in continuing to use a legacy solution that is past its prime, and no longer being actively updated. Hadoop was designed and built before many of the modern data architectures were in use, and fundamentally reflects an out of date architectural model. Current generation solutions include advanced functionality like hybrid and multi-cloud support, cloud-native on premises capability, modern data architecture enablement, data streaming, advanced analytics, multi-layered security, powerful data management, etc. Below is our evaluation of why we think few if any organizations should remain on Hadoop.

"...If you are one of the companies still using Hadoop, you should be asking yourself if that's the right approach given the changes in modern data processing and hybrid computing requirements. We believe that the cons greatly outweigh the pros in continuing to use a legacy solution that is past its prime, and no longer being actively updated....."

Hadoop's Many Challenges, a few of which include:

Designed for "build it yourself" environments

There is no single complete solution Hadoop capability available off-the-shelf. Rather, Hadoop is meant to be a customizable basket of technology that users must configure and deploy. This requires a significant effort by IT staffs that are well versed in its implementation and ongoing support needs. And while there is an ability to obtain a commercial version of Hadoop from some vendors, it simply moves the significant amount of implementation effort required to an "outsourced" resource.

Requires a large number of components that need management and customization

A typical Hadoop installation not only consists of Hadoop core, but also a variety of peripheral components that also need to be configured and managed. In some cases it can have 10-20 added components that need to be attached, customized and managed to deliver needed and expanded functionality. This often includes components like an analytics engine (Spark), event store and stream processing (Kafka), data query (Hive), etc., and represents a very heavy lift for companies who are often resource constrained. The potential of open source's low cost (there is no such thing as a free lunch) for using Hadoop is often overtaken by the sheer effort involved in configuring all the additional open source components needed for a complete solution.

Designed for Very Large File Storage and Batch Processing.

Hadoop was conceived as a large data repository. It was never conceived of being what most modern data lakehouses have become – a repository, often distributed and "streamed" when accessed, for many different data types of varying size that are critically utilized by functional analysis capabilities, either through traditional data analysis tools, or increasingly

through the use of AI.

Security Issues with Java code
There are many examples of the relative lack of security in the Java language with attacks and data breaches happening often. This is further complicated by the relatively customized features implemented by IT designers who may not have placed security front and center, and the need to add-on peripheral code for a complete solution that may also lack effective security. The need for a unified security structure was never really a part of the Hadoop philosophy.

Lack of Inherent Analytics and/or AI capability
As companies expand, both in size and complexity, as well as in data needs from internal and external sources, it's often difficult for them to continue with existing implementations of their legacy data warehousing. Indeed, distributed data processing, and particularly analysis of hybrid data sets from a variety of sources, is not easily accomplished. With so many distributed environments, doing analysis exclusively on a central repository of data is often impractical or impossible. And the lack of modern analytics capability within Hadoop itself leads to component and/or solutions integrations.

Can Hadoop be "wrapped" to reduce its failings?
Simple putting a "wrapper" around a Hadoop installation does not solve a fundamental problem. It does make it easier to manage the installation and the variety of parts and peripheral components. But it does nothing to modernize the overall data accumulation and access tasks necessary to run a modern, cloud first, analytics driven organization that not only accumulates, but consumes data in near real time. The result is that putting a wrapper around Hadoop doesn't solve its underlying deficits.

"...Our analysis shows that while Hadoop was ahead of its time when built, the major challenges it presents makes it difficult for any organization to run a modern fully data driven and hybrid environment. While legacy apps may continue to run in Hadoop, they are not fully compatible with a modern data driven organization, and enterprises continuing to rely on Hadoop run the risk of becoming incapable of utilizing the most advanced capabilities in hybrid data and cloud infrastructure...."

Looking at Cost
The cost of configuring and managing a Hadoop installation, supposedly free as open source software, can actually be several times the cost of utilizing an optimized commercial data lakehouse, particularly as the complexity of data access and analysis increases. In a hybrid cloud world, Hadoop does not play very well without a significant number of add-ons, thus raising complexity and potential points of inefficiencies, or even failure. Failures can lead to tens of thousands of dollars in unexpected expenses, critical delays in data analysis, as well as negatively affecting business operations and relations with customers. Indeed, some estimates show that an open source implementation can be 4X to 7X as expensive to operate and maintain as a commercial solution, and that doesn't even include the potential for producing greater insights that could be highly advantageous to the business operations. Further, with a lack of an integrated security model for Hadoop and its various components, and the potential of an enterprise data breach, implementation can prove very costly indeed. Finally, the cost of hiring, training and retaining staff with specialist skills is not an insignificant issue, and may include an added burden should they choose to unexpectedly leave the company.

The Challenges of Hybrid Data Access, Distribution, and Elimination of Data Silos
It's imperative that enterprises are positioned to make use of a hybrid data cloud in order to achieve maximum efficiency, while also enabling use of the most advanced tools available within an increasingly distributed world of data sets, remote operations, shared data and specialized analysis tools. Data generated at most organizations is becoming increasingly varied and dispersed, as more system and workplace solutions come online. This may often include both internal and external cloud implementations of various mission-critical solutions, together with their own data sets. Yet sharing data from all of the various systems and solutions is critical to achieving the lowest TCO and maximizing operational efficiency. With a majority of companies now looking at cloud migration, the ability to utilize a hybrid data structure that was never envisioned by the creators of Hadoop, is becoming mission critical.

The Need for an Integrated Security Architecture
Hadoop and its many additional components, was not built on a single integrated security model. Hadoop was created using Java which presents many security challenges. Indeed,

adding independent functional components exposes a security challenge that prevents a uniform security model from being fully implemented and creating potential security exposures. With the average data breach costing $9.44M in the US according to the 2022 IBM Data Breach Report, the need for a foundational and complete security architecture across the entire data application is critical.

"…Staying on Hadoop means relying on an old and largely outmoded data environment that prevents the use of the most capable tools and enhanced cloud feature sets, including advanced features offered by the hyperscalers. We believe nearly all implementations would benefit greatly from moving off of Hadoop and onto a more modern solution……"

Modern Hybrid Approaches and Common App Requirements

Most companies are moving to a hybrid cloud model. Private cloud is good for some workloads while public cloud is optimum for others. And in most organizations, there are multiple LOBs that create and access data pertinent to their immediate needs. But data needs to be universally available in a secure and easily transferable way in order to create a corporate wide analysis capability. The key to making this happen is the ability to not necessitate storing all data into one repository, but rather have a platform that can ingest data as needed from whatever location it originates. That's the promise of Hybrid Cloud data streaming capabilities. Companies not deploying such a strategy run the risk of creating a significant number of "siloed" data sets that can't be easily accessed. This creates a roadblock to being a fully data driven organization.

Further, the ability to use a common set of data services inherent in a hybrid data cloud approach enables the organization to write an app that can be used universally across a divergent set of public and private clouds, which often are not fully compatible from an app perspective. This universal translation, together with proven security architecture, removes much of the burden of code incompatibility, especially when using multiple cloud service providers. It also provides the ability to send workloads to the most appropriate compute resource, potentially saving many dollars and time to process. This is not a feature available in an outdated data warehouse model like Hadoop.

Moving Forward

Hadoop was built in the past, not built for the future. What's needed is a modern architecture, fully cloud native and hybrid data enabled, and functioning beyond a simple data warehousing solution. Contrast Hadoop with a product like Cloudera's CDP, which includes:

- Consistent security implemented across all layers of functionality and interactions
- A uniform management console with an integrated view of overall process
- Easy ingress and egress of data
- Streaming from multiple data sources including various cloud instances
- Minimal resource commitment required from IT
- Fully hybrid data and cloud optimized
- Supports nearly any data type, of any size for processing and storage
- Real time or batch processing enabled
- Easily customizable to alleviate the need for specialized IT resources

Bottom Line: Our analysis shows that while Hadoop was ahead of its time when built, the major challenges it presents makes it difficult for any organization to run a modern fully data driven and hybrid environment. While legacy apps may continue to run in Hadoop, they are not fully compatible with a modern data driven organization, and enterprises continuing to rely on Hadoop run the risk of becoming incapable of utilizing the most advanced capabilities in hybrid data and cloud infrastructure. Enterprises must fully assess the challenges in maintaining a Hadoop installation against the costs and potential data insight advantages of moving to a more modern hybrid data architecture. Staying on Hadoop means relying on an old and largely outmoded data environment that prevents the use of the most capable tools and enhanced cloud feature sets, including advanced features offered by the hyperscalers. We believe nearly all implementations would benefit greatly from moving off of Hadoop and onto a more modern solution.

J.Gold Associates, LLC.
6 Valentine Road
Northborough, MA 01532 USA

Phone:
+1-508-393-5294

Web:
www.jgoldassociates.com

Email:
info@jgoldassociates.com

Research, Analysis, Strategy, Insight

J.Gold Associates provides advisory services, syndicated research, strategic consulting and in-context analysis to help clients make important technology choices and enable improved product deployment decisions and go to market strategies.