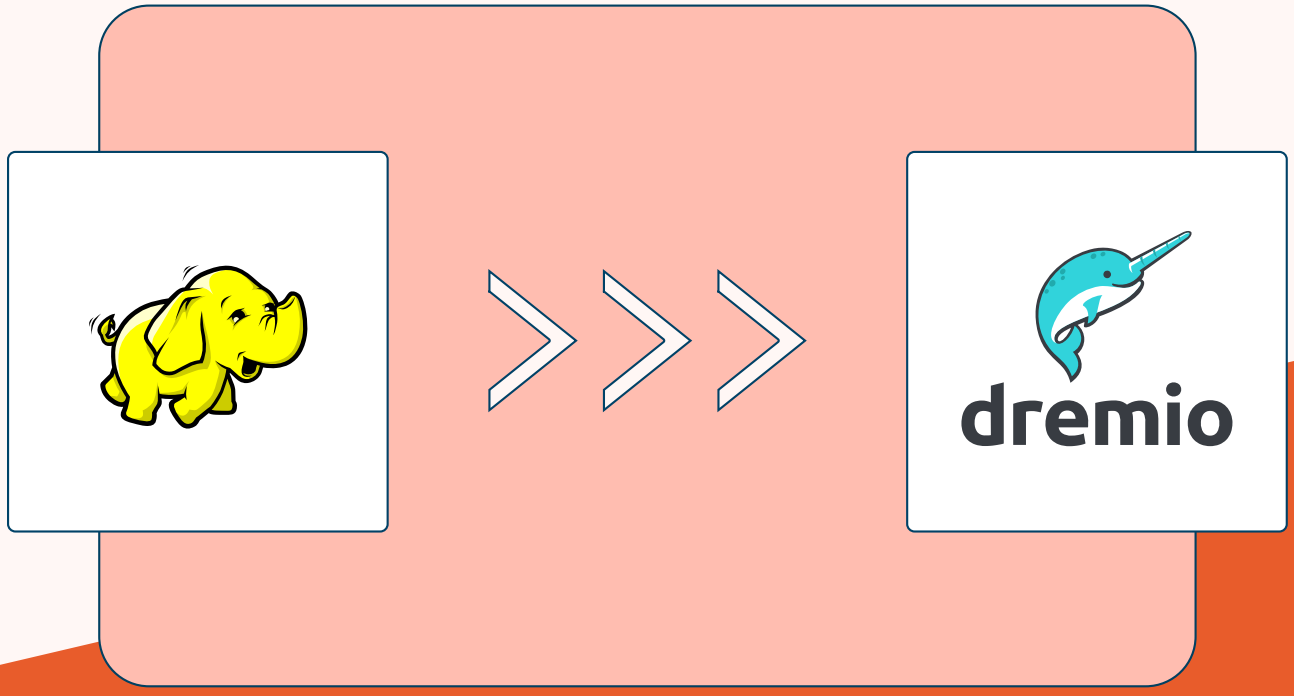
dremio

WHITEPAPER

From Hadoop to Data Lakehouse: A Migration Playbook

Three steps to transform Hadoop into an open architecture for self-service analytics

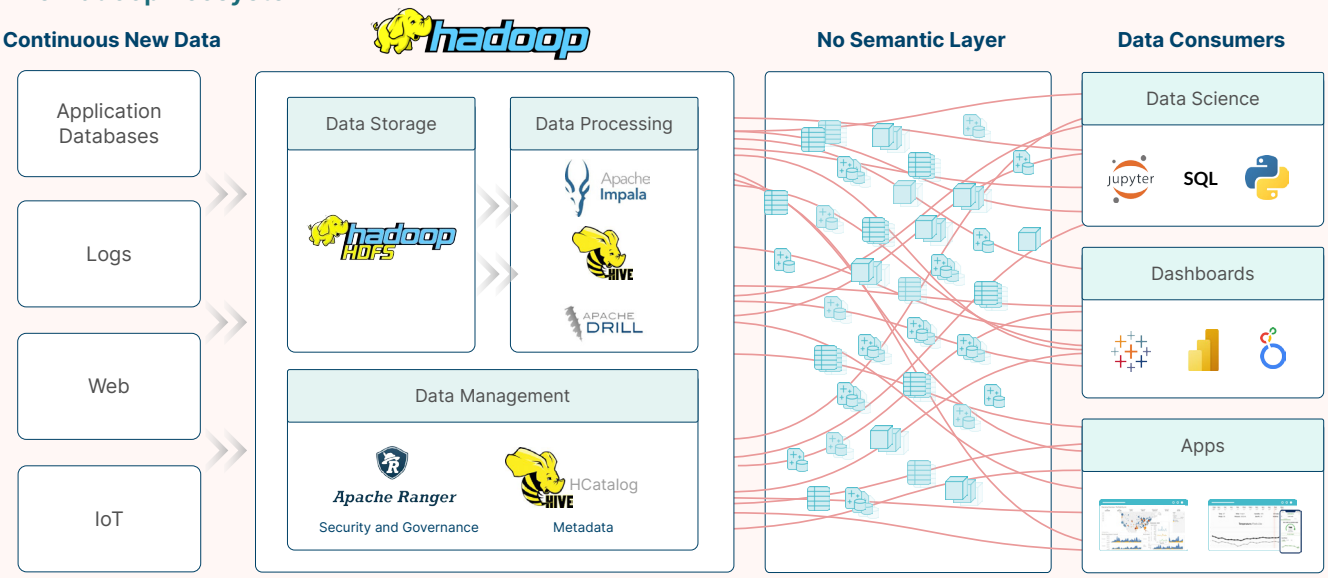
Author: Donald Farmer, Principal at TreeHive Strategy



Challenges

- ✓ Requires deep expertise in the Hadoop ecosystem to maintain
- ✓ High cost of scalability as your data grows
- ✓ Query performance management
- ✓ Difficult to enable governed self-service analytics

The Hadoop Ecosystem



The Hadoop ecosystem as shown in the above diagram comes with a number of challenges associated with various common components including HDFS, Hive, Drill, Impala, Ranger, and Hive metastore.

- At the storage layer, HDFS can become a bottleneck especially as the number of files and directories increases. This can lead to slow read and write operations and ultimately impact the performance of the entire system.
- The Hive query engine is designed for batch processing and is not optimized for interactive queries, which can result in long query execution times.
- Meanwhile, Impala (a SQL query engine, although with a limited SQL implementation) is designed to provide fast interactive queries, it is optimized only for small to medium-sized clusters.
- In contrast, Apache Drill is an open-source distributed SQL query engine, providing high performance over large and complex datasets. It is also highly configurable, but with this comes considerable complexity, making it difficult for users to set up and manage.

"Hadoop is cheap to provision and expensive to run."

Data management features suffer from similar complexities ...

- Apache Ranger is a security management tool that provides centralized administration. It is designed to provide granular access control and auditing but it's difficult to configure. Moreover, it can prove to be resource-intensive in practice, which in some cases can impact the performance of the entire Hadoop cluster.
- Hive metastore is a repository describing Hive tables and partitions. It is somewhat difficult to scale, but more importantly it's metadata support is limited especially for complex data types. Data Lineage is not comprehensively supported and there is limited statistical metadata available without additional analysis.

At one time, the advantages of Hadoop overcame these drawbacks and their hidden costs. But over the years many enterprises have found that Hadoop is cheap to provision and expensive to run. The case for migration is compelling, but the task ahead can appear formidable, especially for businesses that have committed so much to the Hadoop platform. Let's look at some alternatives.