CITO Research

Advancing the craft of technology leadership

The Operational Data Lake:
Your On-Ramp to Big Data

SPONSORED BY

splice MACHINE

CONTENTS

Introduction

Companies are increasingly recognizing the need to integrate big data into their real-time analytics and operations. For many, though, the path to big data is riddled with challenges – both technical and resource-driven. On the other hand, many organizations have operational data stores (ODSs) in place, but these systems, while useful, are expensive to scale. This has led many of those companies to consider upgrading their ODSs. Meanwhile, other organizations are trying to find the right use cases for big data, but may not have the expertise to derive immediate value from implementing Hadoop.

This CITO Research white paper discusses a new concept: that of the operational data lake, and its potential as an on-ramp to big data by upgrading outdated ODSs.

The ODS Is at the Breaking Point

For years, the ODS has been a steady and reliable data tool. An ODS is often used to offload operational reporting from expensive online transaction processing (OLTP) and data warehouse systems, thereby significantly reducing costs and preserving report performance. Similarly, an ODS can prevent real-time reports from slowing down transactions on OLTP systems. An ODS also facilitates real-time reports that draw data from multiple systems. For example, let's say you wanted to run a report that showed customer real-time profitability: an ODS would allow you to pull data from your financial system, CRM system, and supply chain system, supporting a comprehensive view of the business.

When compared with a data warehouse, an ODS offers a real-time view of operational data. While data warehouses keep historical data, an ODS keeps more recent data. Another important use case for an ODS is supporting the ETL (Extract, Transform, Load) pipeline. Companies use ODSs as a more cost-effective platform than data warehouses to perform data transformations and ensure data quality, such as data matching, cleansing, de-duping, and aggregation.

Big data repositories such as Hadoop are causing organizations to question whether they want to keep all of the ODSs they have put in place over the years

Big data repositories such as Hadoop are causing organizations to question whether they want to keep all of the ODSs they have put in place over the years. After all, Hadoop promises many of the same benefits. Hadoop is cost-effective because it uses scale-out technology, which enables companies to spread data across commodity servers. An ODS, on the other hand, uses outdated scale-up technology. Scaling an ODS is prohibitively expensive, requiring more and more specialized hardware to get the required performance. Plus, most scale-up technologies become creaky when they exceed a terabyte of data – which is increasingly common.

Introducing Hadoop for Big Data

To experiment with big data, many companies decide to implement Hadoop. While Hadoop is a great platform for unstructured data, it is not as conducive to structured, relational data. Hadoop uses read-only flat files, which can make it very difficult to replicate the cross-table schema in structured data.

While Hadoop is a great platform for unstructured data, it is not as conducive to structured, relational data

An important use case for Hadoop is ETL. However, since Hadoop flattens structured data into files, it creates additional steps in the ETL process. In addition, with Hadoop being read-only, the ETL pipeline can become fragile and brittle in the face of data quality issues or ETL failures. Without the ability to update data when there is a problem, everything needs to be trashed and restarted, leading to excessive delays and missed ETL update windows.

The Best of Both Worlds: The Operational Data Lake

What if you could upgrade the ODS so you could scale it affordably, while at the same time providing yourself with a platform to experiment with unstructured big data? This is the principle behind the operational data lake, which includes, at its heart, a Hadoop relational database management system (RDBMS).

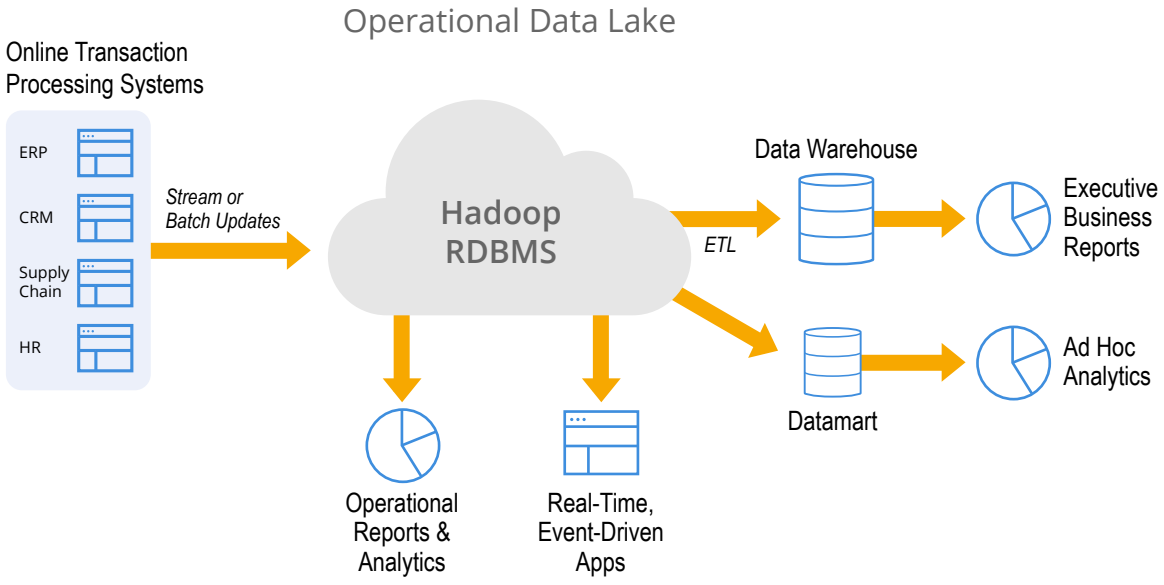

Figure 1. Operational Data Lake Architecture

A data lake is a repository for large quantities of both structured and unstructured data. It allows companies to store data in its native format, while maintaining the integrity of the data and allowing different users to tap into the data in its original form. The operational data lake is the structured part of the data lake, powered by an operational RDBMS built on Hadoop.

Because the operational data lake is built on Hadoop, it offers all the capabilities of Hadoop, enabling organizations to move into big data at their own pace while getting immediate value from operational data stored in Hadoop.

Companies may consider adding a Hadoop-based operational data lake as an ODS replacement or, if they have already implemented Hadoop, consider adding a Hadoop RDBMS to their existing data lake. Here are some details about each of these approaches.

ODS Replacement. For organizations with an ODS in place, a Hadoop-based operational data lake allows companies to:

- Deploy an affordable scale-out architecture
- Leverage existing SQL expertise and applications
- Speed up operational reports and analytics by parallelizing queries
- Augment structured data with semi-structured and unstructured stored in Hadoop

Because operational data stores can be migrated to Hadoop, the operational data lake reduces the cost of using an ODS and offers the opportunity to offload workloads from expensive data warehouses. Companies spend millions of dollars on enterprise data analytics and data warehouses. As much as half of the workloads sitting in one of those warehouses can be handled in a more cost-efficient operational data lake.

Existing Hadoop-Based Data Lakes. For companies that have already made the jump to Hadoop and created a Hadoop-based data lake, adding a Hadoop RDBMS provides the following benefits:

- A native way to store structured, relational data without having to flatten it into read-only Hadoop files (see Figure 2)
- The capability to offload ETL transformations and processing from expensive data warehouses and/or dedicated ETL scale-up platforms (e.g., Informatica, Ab Initio)
- The ability to gracefully recover from data quality issues and ETL failures in seconds with incremental updates and rollback, instead of the hours needed when restarting the ETL process

While unglamorous, ETL processing often represents an expensive, cumbersome process at many companies. An operational data lake can significantly reduce costs and provide a robust ETL processing pipeline that can recover in seconds from data quality issues and ETL failures.

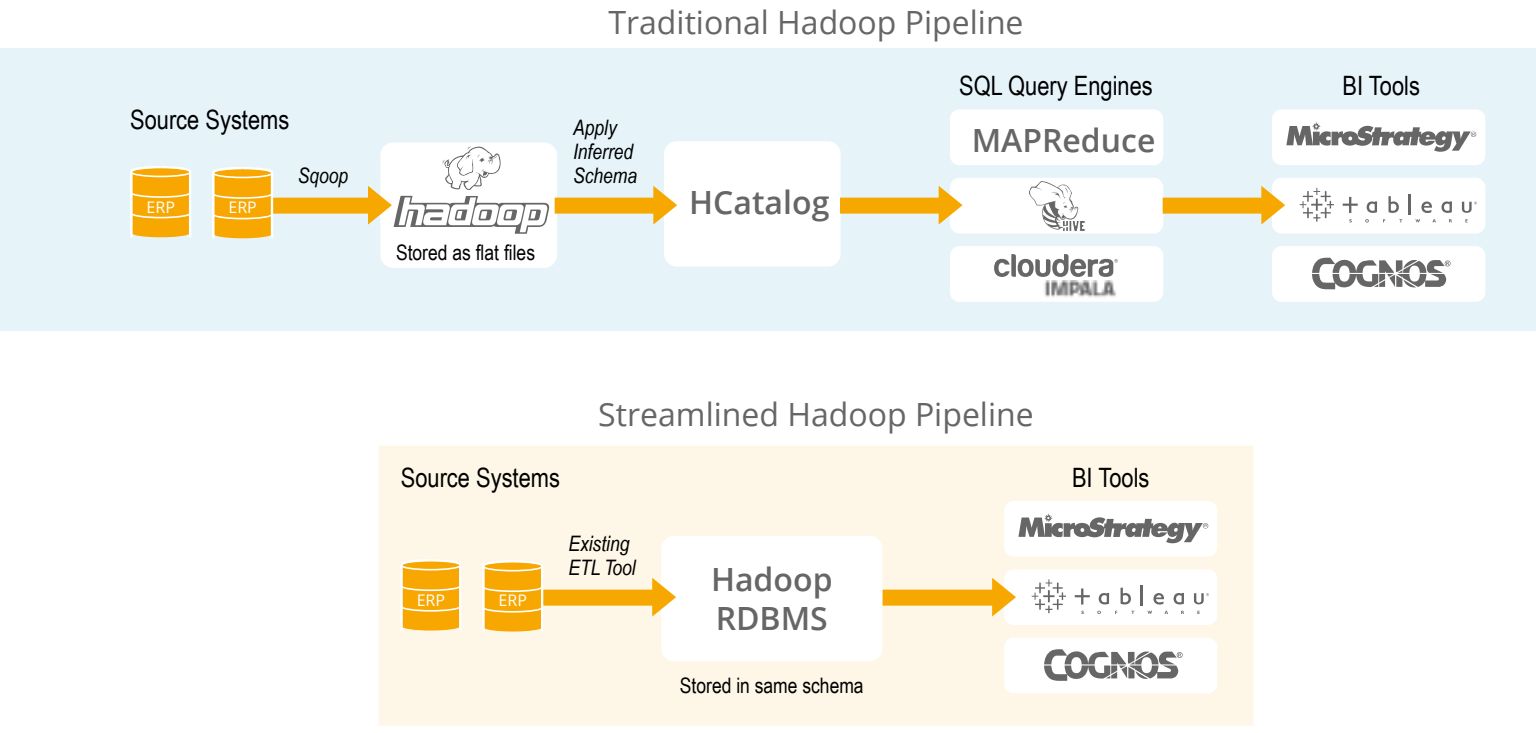

Figure 2. Streamlined Structured Data Pipeline

A data lake is operationalized via a Hadoop RDBMS, where Hadoop handles the scale out, and the RDBMS functionality supports structured data and reliable real-time updates. With this setup, the data lake is never overwhelmed like a traditional ODS. It's nearly bottomless or limitless in its scalability – companies can continue to add as much data as they want because expansion costs so little.

The Operational Data Lake as a Bridge to Big Data and Hadoop

Jumping into Hadoop without a Hadoop RDBMS can leave companies feeling stranded in a sea of big data. An operational data lake is the perfect stepping-stone to big data. Oftentimes, companies start with experimental big data Hadoop clusters in what is essentially technology in search of a use case. Rather than trying to find an initial compelling use for big data, an operational data lake allows companies to expand on what they are already doing with an ODS and build up their big data practice at their own pace. Both are supported on the same commodity hardware.

An operational data lake is the perfect stepping-stone to big data

With the data lake, users can extract structured metadata from unstructured data on a regular basis and store it in the operational data lake for quick and easy querying, thus enabling better real-time data analysis. And, just as importantly, because all data is in a single location, the operational data lake enables easy queries across structured and unstructured data simultaneously.

Finally, unlike native Hadoop, an operational data lake can handle CRUD (create, read, update, delete) operations in a highly concurrent fashion. The system can handle truly structured data in real time, while using transactions to ensure that updates are completed in a reliable manner.

Splice Machine in Action

Based in San Francisco, Splice Machine provides the only Hadoop RDBMS. It is designed to scale real-time applications using commodity hardware without application rewrites. Splice Machine's goal is to provide companies with an ACID-compliant, massively scalable database for applications that doesn't require you to compromise SQL support, secondary indexes, joins and transactions.

Splice Machine optimizes the operational data lake by marrying two proven technology stacks: Apache Derby and HBase/Hadoop. With over 15 years of development, Apache Derby is a popular, Java-based ANSI-SQL database, while HBase enables real-time, incremental writes on top of the immutable Hadoop file system. Splice Machine does not modify HBase; it can be used with any standard Hadoop distribution, such as Cloudera, Hortonworks, and MapR.

Splice Machine's Hadoop RDBMS offers exceptional performance while reducing the cost of traditional RDBMSs like Oracle by 75-80%. By parallelizing query execution, Splice Machine can increase query performance by 5-10 times compared to other RDBMSs.

Splice Machine Busts the Query Performance Blues

Marketing services company Harte Hanks found that its queries were slowing to a crawl, taking a half an hour to complete in some cases. Given the company's prediction that its data would grow by 30 to 50%, query performance would only get worse.

Harte Hanks replaced its Oracle RAC databases with Splice Machine. Rob Fuller, the company's Managing Director of Product Innovation, saw queries that took 183 seconds on Oracle complete in 20 seconds on Splice Machine. Another query with a complex set of joins completed in 32 minutes on Oracle and just 9 minutes on Splice Machine.

Harte Hanks has experienced a 3-to-7 fold increase in query speeds at a cost that is 75% less than its Oracle implementation.

Conclusion

Big data offers valuable insights that companies need to succeed in today's increasingly competitive marketplace. However, many companies do not know where to start and don't want have an unending big data "science project." Upgrading ODSs with obsolete technology is an excellent place to start. An operational data lake is the next-generation ODS. At CITO Research we believe that the operational data lake enabled via a Hadoop RDBMS is a great choice for companies that want to leverage their existing skills while ramping up their big data use cases.

With an operational data lake, companies can significantly reduce costs by offloading operational reports and ETL processing from expensive OLTP systems and data warehouses to a scale-out architecture based on commodity hardware. It also enables companies to experience faster processing times, improve user satisfaction, and access more data.

Learn more about Splice Machine

This paper was created by CITO Research and sponsored by Splice Machine

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com