DataInformed
Big Data and Analytics in the Enterprise

The Guide to Real-Time Hadoop

Table of Contents:

A Hadoop Primer: What It Is and How It's Used

by Scott Etkin

With all the hype around big data, many organizations find themselves wrestling with questions about Hadoop and what value it delivers. This primer can help get you up to speed on Hadoop: what it is, what it does, and why to use it.

With data production accelerating to unprecedented rates, many organizations have turned to Hadoop as an inexpensive way to store and process that data. But those new to Hadoop often find themselves confronting a technology as inscrutable as its name. What is Hadoop? What does it do? What's with the elephant?

To put it simply, Hadoop is an open-source distributed platform for storing and processing large data sets. Inspired by research papers published by Google, Doug Cutting and Mike Cafarella created Hadoop in 2005, when Cutting was working at Yahoo on its search platform.

Named after a toy elephant owned by Cutting's son, Hadoop helps control big data processing costs by distributing computing across commodity servers instead of using expensive, specialized servers. Commodity hardware refers to machines that are inexpensive, already available or easily obtainable, and interchangeable with similar hardware. Commodity computing is seen as a more scalable solution because purchasing additional high-cost, high-performance servers to keep up with the growth of big data quickly becomes prohibitively expensive.

After creating Hadoop, Yahoo turned it over to the Apache Software Foundation, where it is maintained as an open-source project with a global community of users and contributors. The Apache Hadoop framework is made up of the following modules:

- Hadoop Distributed File System (HDFS), a Java-based distributed file system designed to store data on commodity servers. HDFS stores files by dividing them into smaller blocks and replicating them on three or more servers.
- MapReduce, a batch programming model that uses parallel processing — basically, using more than one CPU to execute a task simultaneously — to make processing big data faster, cheaper, and more manageable.
- Hadoop YARN, a resource-management and scheduling platform. YARN removes the resource management and scheduling responsibilities from MapReduce, optimizing cluster utilization and allowing MapReduce to focus on data processing.
- Hadoop Common, which consists of the libraries and utilities required by other Hadoop modules.

Beyond these core components, Hadoop includes an entire ecosystem of technologies based on HDFS, such Apache HBase (a key-value data store inspired by Google's Big Table), Apache Hive (a SQL-based, analytic query engine), Apache Pig (procedural language), Apache Spark (a fast, in-memory engine for large-scale data processing), and even commercial technologies such as Splice Machine (a Hadoop RDBMS with joins and transactions).

A typical Hadoop environment comprises a master node and several worker nodes. Most Hadoop deployments consist of several master node instances to mitigate the risk of a single point of failure. A Hadoop environment can include hundreds or even thousands of worker nodes.

How It Works

Hadoop differs from a traditional relational database in that it is not, strictly speaking, a database but rather a storage and batch data processing system. In a traditional relational database, data queries are conducted using Structured Query Language (SQL) queries. Hadoop, on the other hand, originally did not use SQL or Not Only SQL (NoSQL) queries, but required Java MapReduce programs for data processing.

MapReduce speeds data processing by bringing the processing software to the data as opposed to moving massive amounts of data to the processing software, as is done in traditional data processing architectures. Data processing, therefore, is distributed, or "mapped," to each node.

MapReduce, in name and function, is a combination of two processing steps: Map and Reduce.

In the Map step, records from the data source are split into bundles for each Map server and are fed into the map() function on that server. Each map function produces a list of results. Each result set is then sent to one reduce server. In the Reduce step, each Reduce server runs a reduce() function over the results lists sent from a subset of map nodes. Then the reducers combine the results of the reduce() runs into a final result.

Using MapReduce is more complex than using a standard database query because it requires the abstraction of tasks into map() and reduce() functions in Java, but tools like Hive can help reduce that complexity by converting a SQL-like query language into MapReduce jobs.

How It's Used

Use cases for Hadoop are as varied as the organizations using Hadoop. Companies have used Hadoop to analyze customer behavior on their websites, process call center activity, and mine social media data for sentiment analysis about themselves, their products, and their competition. Based on this data, companies can make decisions in real time to understand customer needs, mitigate problems, and ultimately gain a competitive advantage.

Hadoop has been used to bolster data security by being applied to server-log analysis and processing machine data that can be used to identify malware and cyber-attack patterns. Banks use it to store customer transaction data and search for anomalies that could identify fraud.

Other organizations might simply use Hadoop to reduce their data storage costs, as commodity hardware is much less expensive than large, specialized servers. Hadoop can be used to cut storage costs and speed up analysis time in just about any data-rich environment.

Hadoop differs from a traditional relational database in that it is not, strictly speaking, a database but rather a storage and batch data processing system.

As mentioned above, Hadoop is an open-source platform. But using the free, open-source version can be a challenge for anyone who isn't a Java programmer. There are commercial distributions of Hadoop (e.g., Cloudera, Hortonworks, MapR) that add functionality and improve reliability. There are also Hadoop-based tools that blur the once-sharp line between Hadoop and traditional RDBMS by providing Hadoop with the capabilities of a RDBMS (e.g., Splice Machine) to run business-critical operational applications.

Hadoop's ability to store, process, and analyze large data sets in a cost-effective manner has made it the de facto architecture for today's big data environments.

Scott Etkin is the managing editor of Data Informed. Email him at Scott.Etkin@wispubs.com. Follow him on Twitter: @Scott_WIS.

How Do I Handle All of That Data? Scale Up or Scale Out?

By Joshua Whitney Allen

As data rapidly accumulates from an ever-growing number of sources, many organizations face the question of how best to manage that data. Is scale up or scale out the best solution for your organization's needs?

Some predictions about the future of big data may seem far-fetched, yet the reality is that big data is already here. Every social media post, every click on the Internet, every sensor measurement from hundreds of millions of devices — they're all piling up. The numbers are staggering: IDC forecasts the scale of all this information at 40,000 exabytes by 2020. To illustrate a volume of data that gargantuan, that's more than 33 times the data stored in the world today. If printed, this would cover the planet in a layer 52 books deep.

Throughout the global economy, this tsunami of data is crushing IT budgets. Gartner projected a global overall IT spending increase of only 3.8 percent for 2014, while information volumes are rising worldwide more than 59 percent each year. In response, companies everywhere are looking to store their data in more cost-effective ways.

From a technology standpoint, IT managers can choose between scale-up or scale-out architectures to handle the deluge of data. Scalability refers to how a system accommodates additional data and workloads. Scaling up involves adding more resources like CPU cores, memory, disks, and network bandwidth to a single server. Scaling out involves adding more servers that cooperate simultaneously.

With data ever expanding, which scaling strategy meets the processing time, updating, maintenance, and fiscal needs of modern IT departments?

There is no shortage of opinions, with experts endorsing different approaches to different situations. Researchers in the UK argued that a scale-up approach worked best for jobs smaller than 100 GB of data. "Clearly there is a job size beyond which scale-out becomes a better option," the writers asserted. "This cross-over point is job specific."

Writing about scale-out designs, ExaGrid Systems' Marc Crespi explained why scale-out performs better at larger scales: "If the workload is quadrupled, the processing power of the architecture is also quadrupled. And there is no 'maximum capacity.' While vendors may limit how many devices can coexist in a singly managed system, there is never the need for a forklift upgrade as devices can continue to be added individually, even if that means starting a 'new system.' "

Because scale-up costs rise significantly faster than performance, a consensus is emerging that scale-out technologies are becoming the future of databases and storage

systems for businesses experiencing rapid data growth. These technologies are better adept at handling massive data volumes, high ingest velocity, and the flexibility of disparate data sources.

Scaling out demands that IT staff adapt characteristics of the solution to a wide array of conditions. On one hand, NoSQL solutions are best used for relatively simple web applications that do not have data dependencies between multiple rows or documents. For instance, Shutterfly, an online photo-sharing company, replaced Oracle with popular NoSQL database MongoDB to more flexibly and cost-effectively manage its metadata storage. With MongoDB, Shutterfly achieved increased agility and 9-times faster performance through parallelized queries.

Scale-out technologies are becoming the future of databases and storage systems for businesses experiencing rapid data growth.

On the other hand, a Hadoop RDBMS like Splice Machine is better suited for existing SQL applications that need reliable updates across multiple rows within a transactional context. Harte Hanks, a marketing services company, also replaced Oracle, with Splice Machine, to support its business-critical applications. They see 7-times faster queries and a 75 percent reduction in overall total cost of ownership as a result.

If the decision whether to scale up or out ultimately comes down to price/performance, then scaling out on commodity hardware by investing either in NoSQL solutions or a Hadoop RDBMS is an affordable yet innovative way to smoothly ride the wave of big data without busting your budget.

Joshua Whitney Allen has been writing for fifteen years. He has contributed articles on technology, human rights, politics, environmental affairs, and society to several publications throughout the United States.

Power a 360-Degree Customer View with a Unified Customer Profile

by Data Informed staff

With a proliferation of web, social, and mobile channels, companies know that they need to deliver true omni-channel marketing. Splice Machine CEO Monte Zweben discusses the advantages of a Unified Customer Profile and how to affordably create one with big data technology.

***Monte Zweben**, CEO, Splice Machine*

It would be ideal if businesses could know their customers better than they know themselves, but creating a 360-degree customer view is easier said than done. Collecting data about your customers' wants and needs is only one part of the puzzle, which remains incomplete until you can turn that data into real-time insights. Customer data that is trapped in silos or that can't be integrated with data collected from other first and third-party sources provides about as much value as an unlit candle.

With a Unified Customer Profile that eliminates those data silos, businesses can use that omni-channel view of the customer to deliver dynamic, more targeted marketing campaigns.

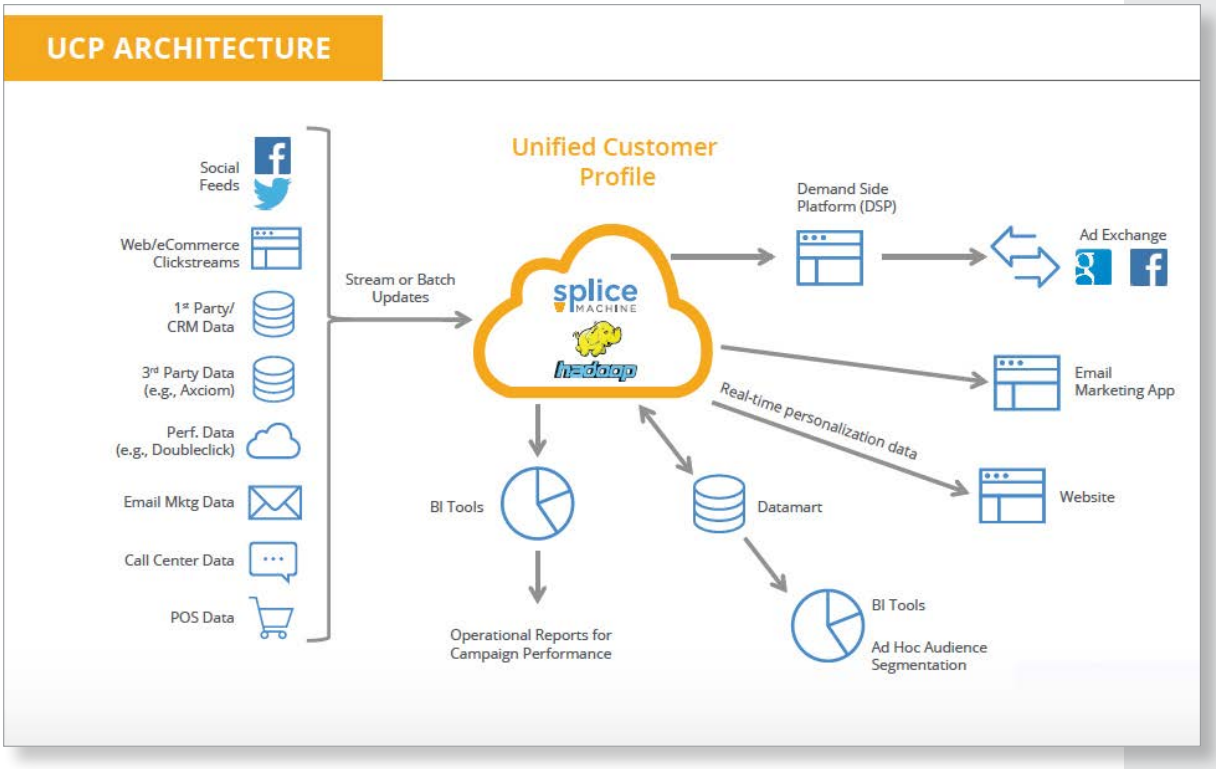
Splice Machine CEO Monte Zweben fielded questions about how to build a Unified Customer Profile and the business advantages that it can provide.

Why should companies be building Unified Customer Profiles? What are the most important business benefits?

Monte Zweben: A Unified Customer Profile (UCP) is a platform that aggregates real-time data across all sources to create a 360-degree view of the customer. A UCP provides deep insight into a customer, enabling companies to deliver more personalized digital marketing campaigns that can be truly coordinated across channels.

What data types/sources can be incorporated to build a UCP and gain a 360-degree view of your customer?

Zweben: A UCP can incorporate all online and offline data from sources, such as point-of-sale transactions, ecommerce clickstreams, call centers, social media feeds, emails, and web logs. UCPs act as data hubs, pulling in both first party and third party customer data to better inform businesses of customer behavior, tailor marketing campaigns, and increase ROI.



What are the technology challenges to building UCPs?

Zweben: Creating a 360-degree view of the customer is by no means a novel concept, but not until recently has technology evolved to make this possible. IDC projected that data currently doubles every two years, so it is getting increasingly more difficult to affordably ingest these massive volumes from all these disparate sources and put the pieces of the puzzle together.

On top of the rising costs of maintaining and analyzing more and more data, it's been challenging to integrate all the data together, given that most companies have it locked up in various silos. This hinders a cohesive UCP that is needed to respond to consumer behavior in real time.

Can you describe Splice Machine's platform for building UCPs?

Zweben: Splice Machine offers the only Hadoop RDBMS, which powers real-time applications overwhelmed by data growth. At a fraction of the cost of traditional databases, Splice Machine is an affordable scale-out alternative that can tackle large data sets and integrate with industry-standard tools for business intelligence and marketing automation.

By utilizing the flexible scalability of Hadoop, Splice Machine can build a UCP that can seamlessly collect and coordinate data across channels, thereby making it easier to run more hyper-personalized campaigns in less time. Our customers in digital marketing have reduced costs by 75 percent and increased performance by 5x-10x.

How does scale-out technology enhance the building of UCPs?

Zweben: Scale-out technology provides an affordable alternative to expensive proprietary scale-up systems. The future of databases in digital marketing is scale-out, because it accomplishes the oft-heard three V's of big data: Scale-out databases can retain massive data Volumes forever, analyze data from a Variety of structures, and achieve real-time personalization with rapid Velocity. Scale-out technology essentially future-proofs your applications, since you don't have to continually invest in costly, disruptive replacements of your existing database hardware.

What is the importance of real-time in building and leveraging UCPs?

Zweben: Digital marketers are discovering more business value from data captured in the last few minutes than historical reports about the last few years. Creating dynamic, personalized content is essential for a UCP, because it has been proven to deepen customer relationships and increase marketing ROI.

HubSpot reported that 60 percent of marketers struggle to personalize content in real time, and the top two reasons were due to the complexity of IT infrastructure and the lack of access to real-time data. Splice Machine eliminates these issues by creating a UCP with that vital 360-degree view of the customer throughout the entire buying process.

What barriers exist for companies looking to build UCPs?

Zweben: There are many databases out there in the market, but the challenge of selecting the right one to build a UCP comes down to performance and integration. Companies must be able to scale dramatically to tackle big data workloads without disrupting their existing applications and tools. Harte Hanks, a leading marketing services provider, replaced Oracle with Splice Machine because we allowed them to manage dozens of terabytes while connecting seamlessly to their campaign management and BI platforms. They now manage an optimized UCP with personalized campaign execution and cross-channel analytics at a quarter of the cost of Oracle. And with a 10x price/performance improvement, they are confident that they can provide a 360-degree customer view to their own clients.

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).

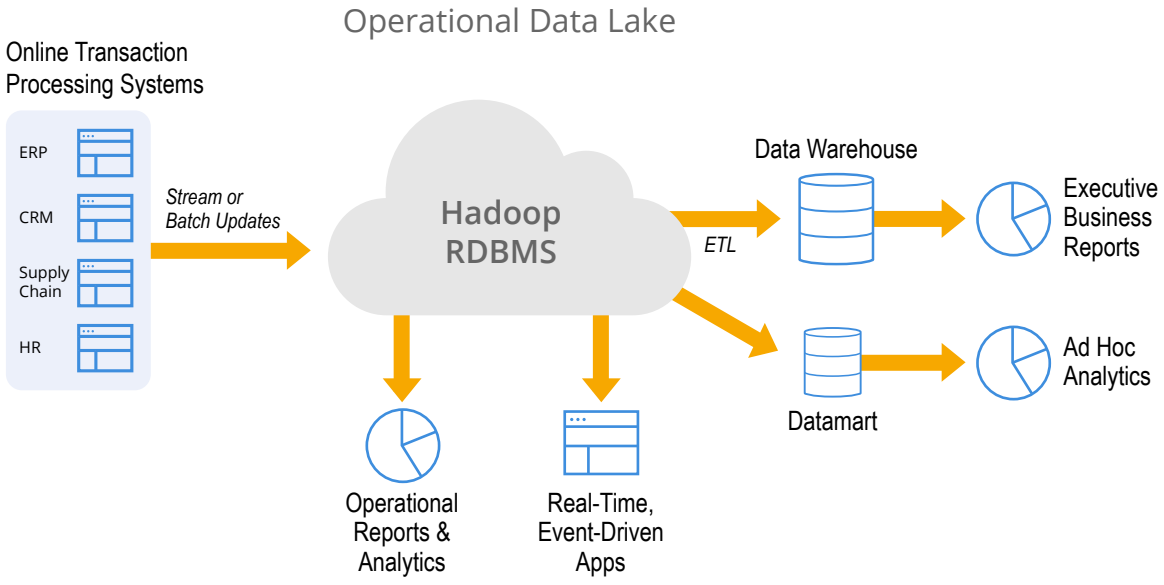


Figure 1. Operational Data Lake Architecture

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.

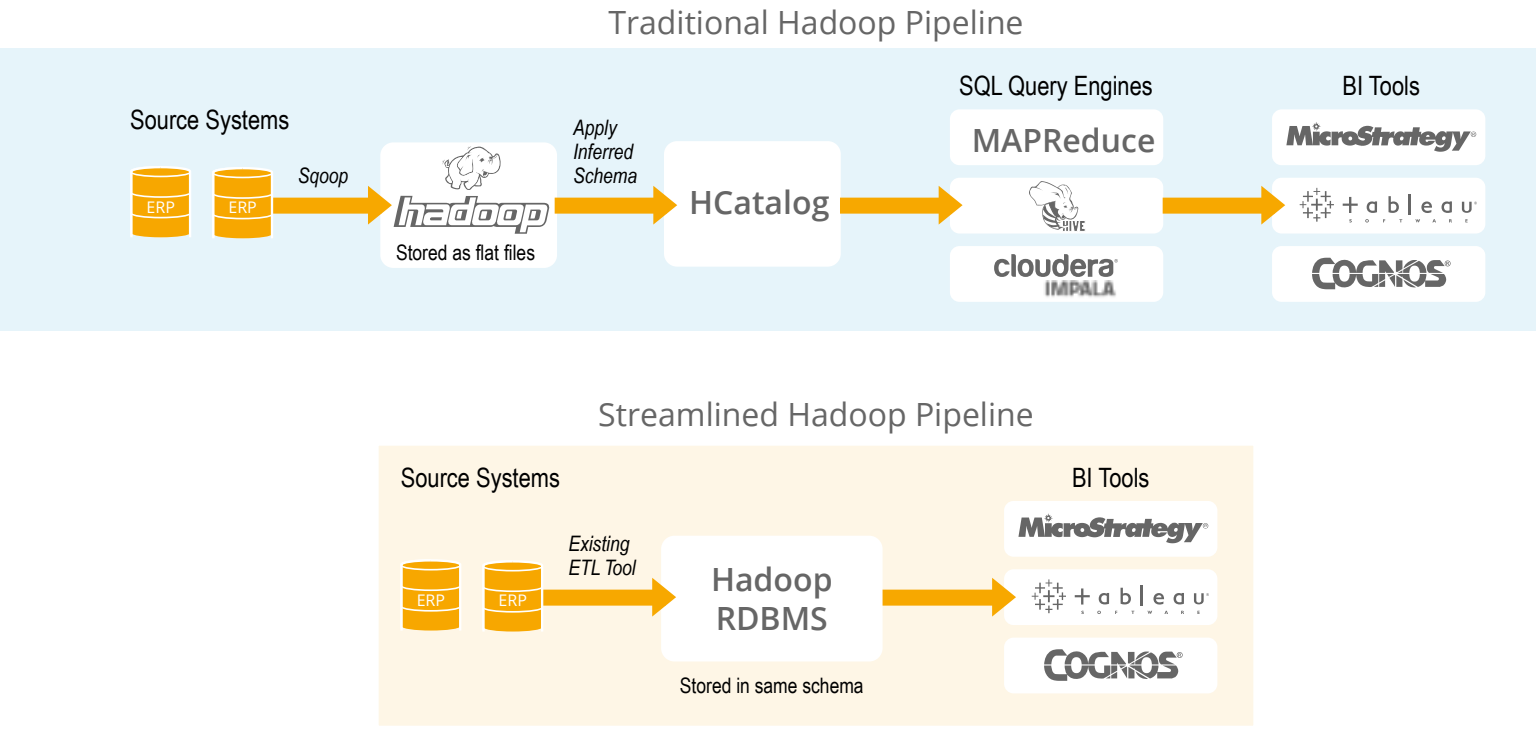


Figure 2. Streamlined Structured Data Pipeline

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

CHECK OUT DATA INFORMED

Find other articles like these and more at *Data Informed:*

www.data-informed.com

Data Informed gives decision makers perspective on how they can apply big data concepts and technologies to their business needs. With original insight, ideas, and advice, we also explain the potential risks and benefits of introducing new data technology into existing data systems.

Follow us on Twitter, @data_informed

DataInformed
Big Data and Analytics in the Enterprise

Data Informed is the leading resource for business and IT professionals looking for expert insight and best practices to plan and implement their data analytics and management strategies. *Data Informed* is an online publication produced by Wellesley Information Services, a publishing and training organization that supports business and IT professionals worldwide.