5 Steps To OFFLOAD YOUR DATA WAREHOUSE WITH HADOOP

A QUICK START GUIDE

Free Up Your Data Warehouse... and Budget

syncsort

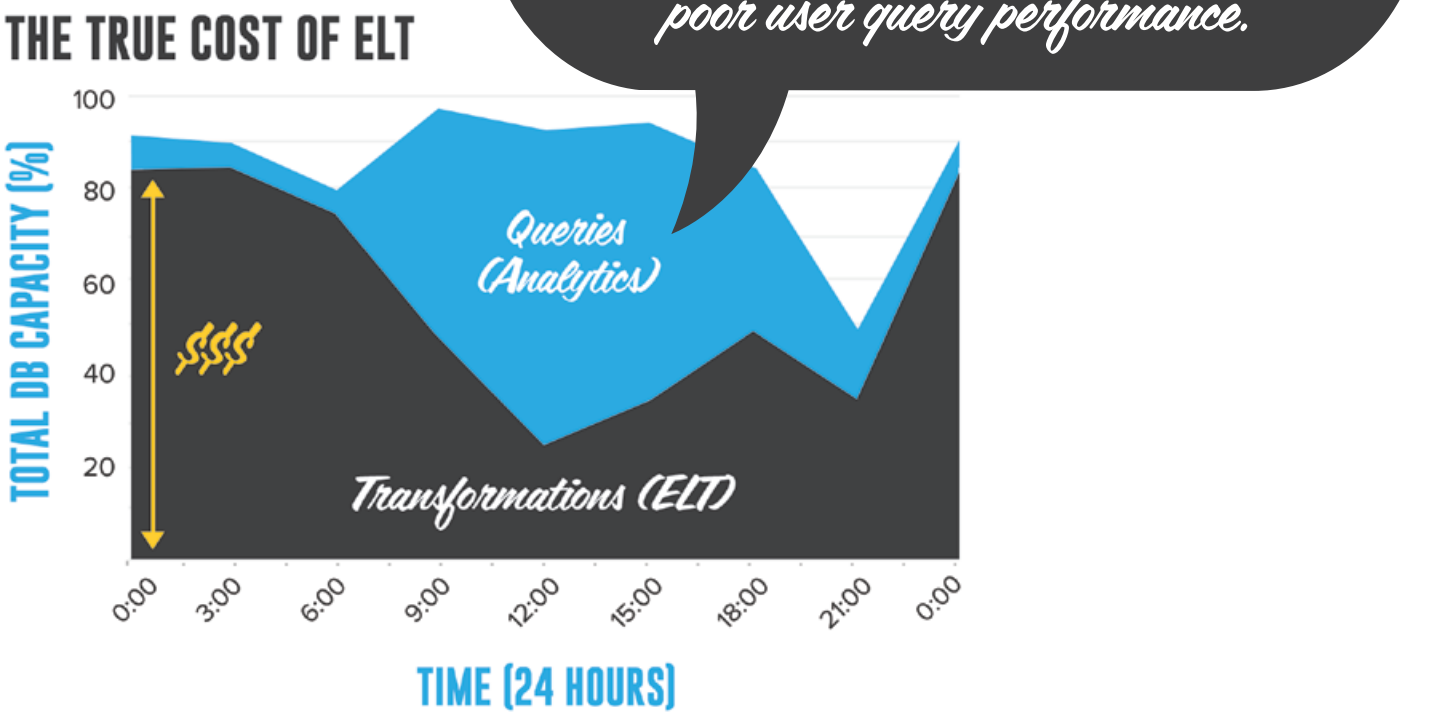
TABLE OF CONTENTS

How did we get here? A brief history of ETL.

They say knowledge is power. But if data generates information which generates knowledge, then isn't data really power? That's why organizations are placing an ever-increasing focus on data as a means to enable better strategic business decisions—but at what cost?

Today, the universe of data measures in the trillions of gigabytes. To unlock its power, data integration and, more specifically, ETL (Extract, Transform, Load) holds the key. ETL is the process by which raw data is moved from source systems, manipulated into a consumable format, and loaded into a target system for performing advanced analytics, analysis, and reporting. For years, organizations have struggled to scale traditional ETL architectures to keep up with the three V's of Big Data – high-volume, high-velocity, and high-variety of data assets. Unable to keep pace, data integration platforms forced IT to push the transformations down to the data warehouse, creating a shift from ETL to ELT. But this proved to be a costly and inefficient approach.

Data integration now consumes up to 80% of database capacity, resulting in unsustainable spending, ongoing tuning and maintenance efforts, as well as poor user query performance.

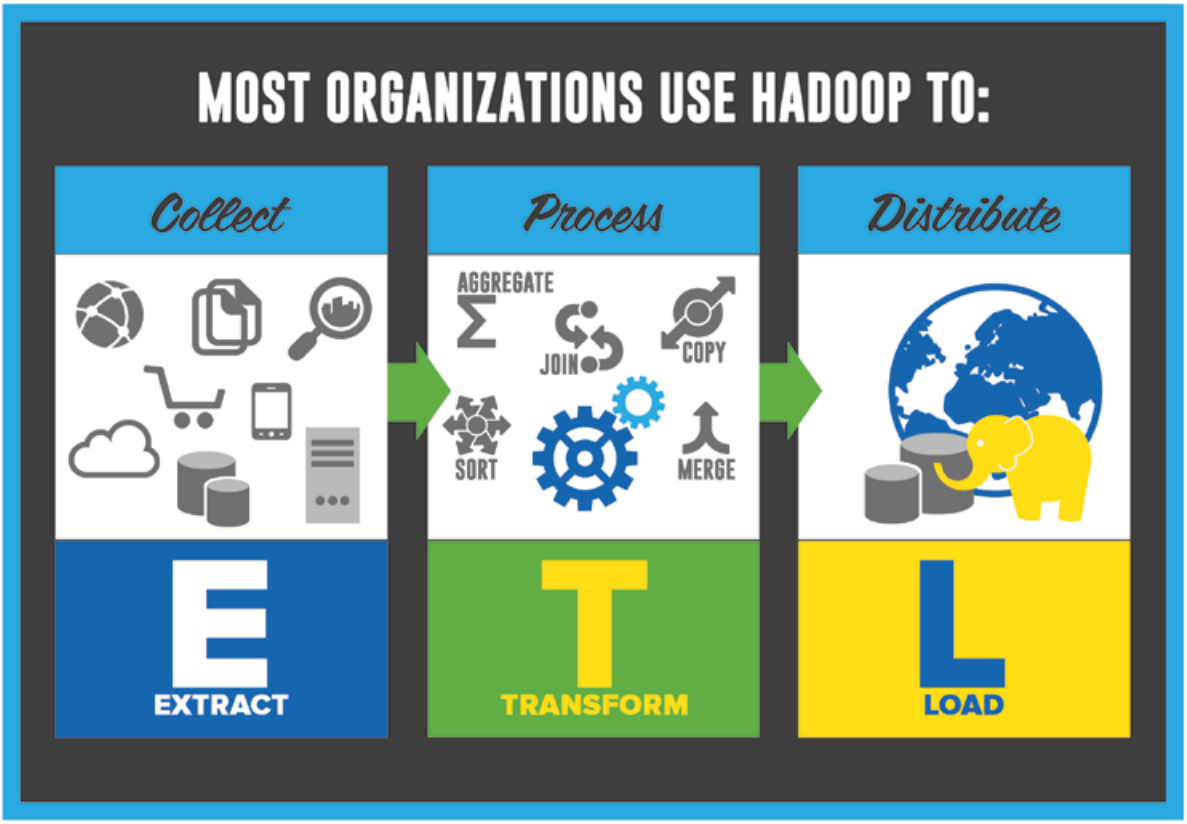


70% OF ALL DATA WAREHOUSES ARE PERFORMANCE AND CAPACITY CONSTRAINED

according to Gartner

It's no surprise then that most organizations cite total cost of ownership as the #1 challenge they face with their data integration tools.

Meanwhile, data volumes continue to grow. In fact, IDC projects data volumes to double every two years. With no end in sight to the digital explosion, organizations are more recently looking at Hadoop to collect, process, and distribute the ever-expanding data avalanche. By shifting ETL to Hadoop – offloading heavy transformations from the enterprise data warehouse (EDW) – organizations are finding they can dramatically reduce costs and free up database capacity for faster analytics and end user queries.



The Hadoop Opportunity

Hadoop has quickly become the new operating system for managing Big Data with massive horizontal scalability along with system-level services that allow developers to create Big Data applications at a highly disruptive cost.

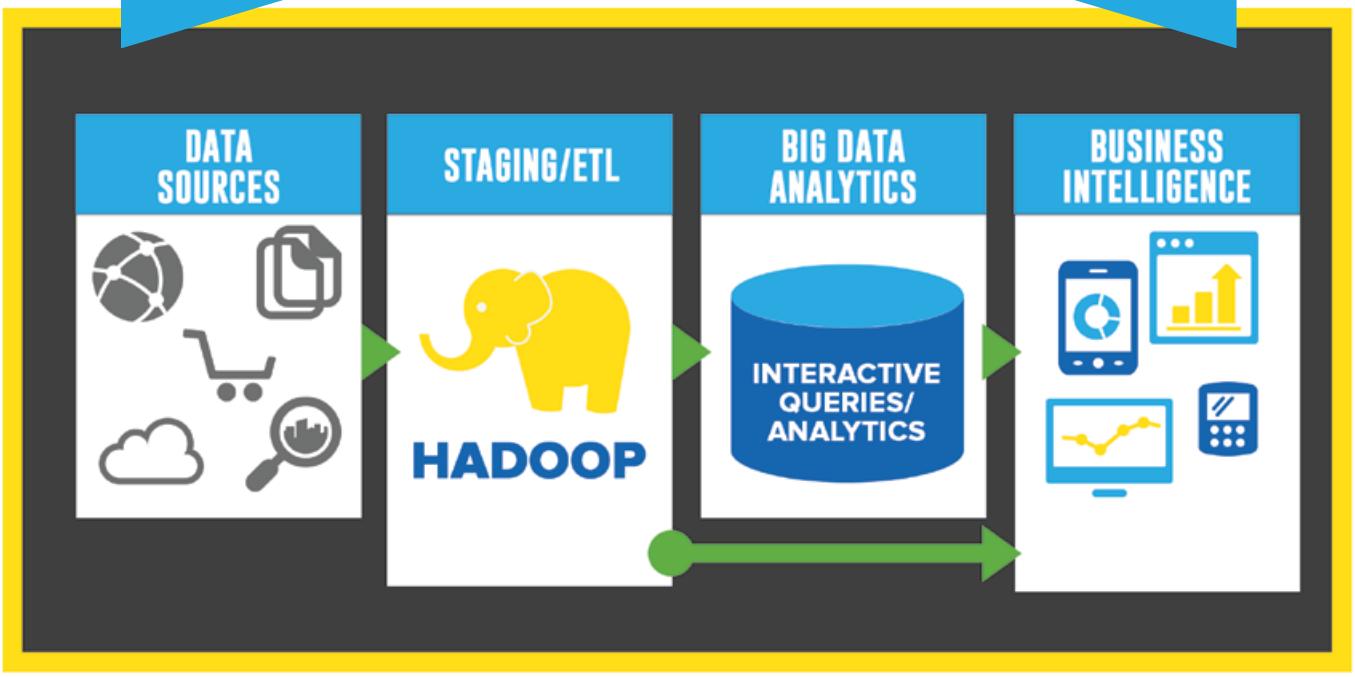
Estimates from multiple sources indicate managing data in Hadoop can range from $250 to $2,000 per terabyte of data, compared to $20,000 to $100,000 per terabyte for high-end data warehouses.

Capitalizing on the efficiency and power of Hadoop, many organizations are embarking on Hadoop ETL initiatives, but Hadoop is not a complete ETL solution. While Hadoop offers powerful utilities and virtually unlimited horizontal scalability, it does not provide the complete set of functionality users need for enterprise ETL. In most cases, these gaps must be filled through complex manual coding, slowing Hadoop adoption and frustrating organizations eager to deliver results.

But there is a way to combine the benefits of purpose-built, high-performance ETL with Hadoop, freeing up the EDW while gaining all the benefits of a complete ETL solution.

Syncsort Hadoop ETL solutions close the gaps between Hadoop and enterprise ETL, turning Hadoop into a more robust and feature-rich ETL solution. Syncsort's high-performance ETL software enables users to maximize the benefits of MapReduce without compromising on capabilities and ease of use of conventional ETL tools. With Syncsort Hadoop ETL solutions, organizations can unleash Hadoop's full potential, leveraging the only architecture that runs ETL processes natively within Hadoop.

A NEW ARCHITECTURE FOR BIG DATA ANALYTICS



This guide will provide five steps to help you get started with offloading your EDW to Hadoop, overcome some of the biggest challenges, and arm you with some best practices to accelerate data integration efforts. Regardless of the tool you choose, these steps will help you free up your EDW to do the work it was meant to do in the first place – provide the insights you need through high-performance analytics and fast user queries.

Step One

UNDERSTAND & DEFINE BUSINESS OBJECTIVES

Most organizations have spent considerable time and money building their existing data integration infrastructure. But increasing demands for information have proven to be too much for traditional architectures to handle on a number of fronts:

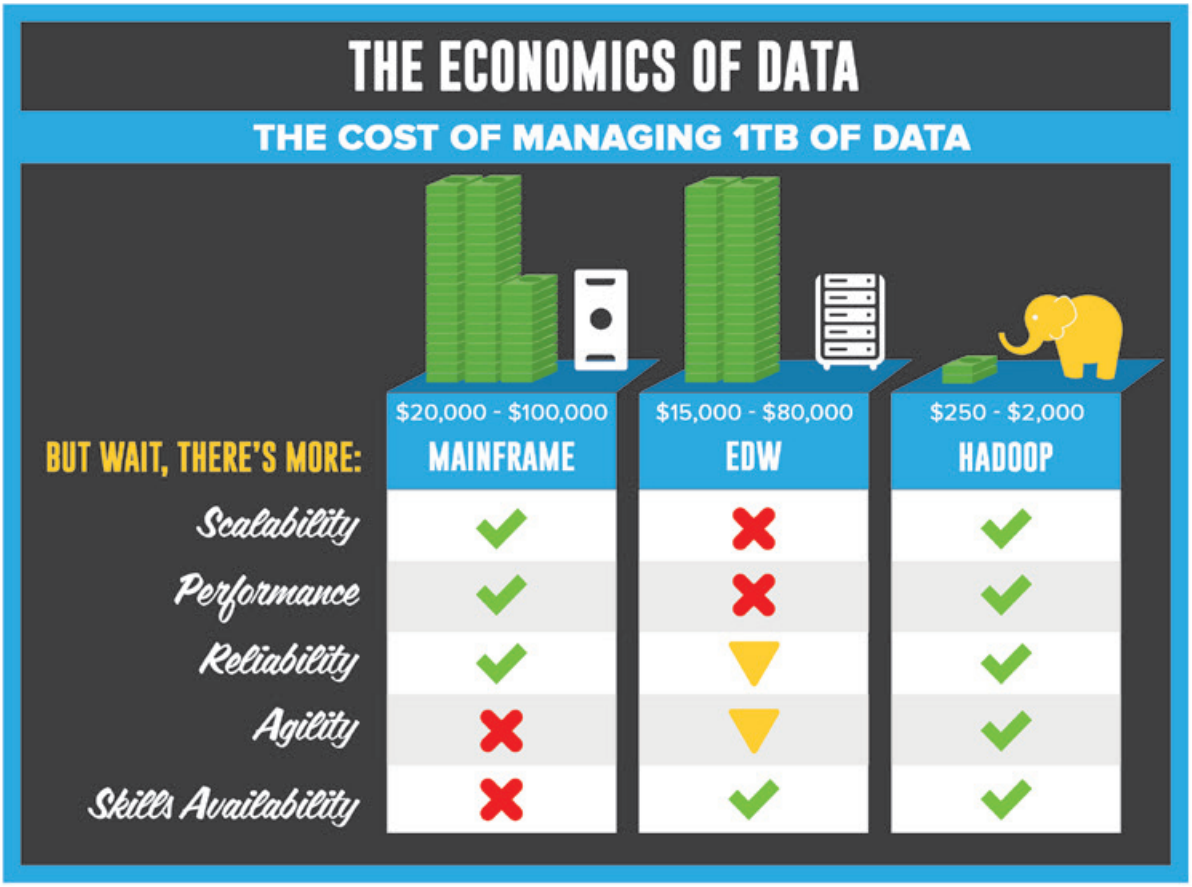
- **Costs:**
  Relying on the EDW for heavy data transformations results in unsustainable costs and complexity. With ELT processes driving anywhere from 40% to 80% of database workloads, it isn't unusual for organizations to spend upwards of $300K per year on additional capacity... just to keep the lights on!

- **Data volumes:**
  Connecting to and managing growing volumes and sources of data in order to make the best business decisions and discover new business opportunities is no longer possible with traditional technologies and manual approaches.

- **Service level agreements (SLAs):**
  Unable to meet SLAs with existing infrastructure, user frustration is on the rise with bottlenecks that turn minutes and hours into days and weeks when waiting for reports.

Whatever your primary objectives may be – reducing costs, leveraging more data, or meeting SLAs – many organizations are forced to look for an alternative approach.

Realizing the critical value of data to sustain competiveness, Hadoop has become the framework of choice to reveal answers previously unknowable, accelerate time-to-insight, and reduce the overall costs of managing data.



By shifting heavy ETL workloads out of the EDW and into Hadoop, organizations can quickly realize significant value including:

- Shortened batch windows
- Fresher, more relevant data
- Faster time-to-insight
- Faster database user query performance
- Defer hundreds of thousands of dollars a year on additional EDW costs

However it's important to remember that Hadoop was not designed as an ETL tool but as an operating system that, with the right tools, enables you to harness the power of Big Data. It isn't realistic to expect it to do everything a high-performance ETL solution can.

There are three main challenges organizations encounter when relying solely on Hadoop for ETL. First, skilled Hadoop programmers are hard to find and costly. Detailed knowledge of Pig, Hive, and Sqoop is essential for creating ETL jobs in MapReduce, yet most organizations don't have that expertise in-house.

Second, the prospect of giving up “enterprise ETL” functionality such as point-and-click interfaces, metadata, reusability, and connectivity jeopardizes gains in productivity. And a return to complex coding means a return to the long development cycles, as well as maintenance and reuse headaches organizations are trying to cure. Third, using the wrong ETL approach with Hadoop can hurt performance by introducing additional overhead. Moreover, while Hadoop excels at horizontal scalability, spreading processing and data across many nodes, there are opportunities to optimize the ETL process to maximize efficiency of each node.

BEST PRACTICES

To ensure delivery against business objectives look for an approach that offers:

- ***A GUI*** with point-and-click interfaces to help keep costs down by leveraging the skill sets of your existing ETL developers
- ***Metadata*** for faster development and greater reusability, boosting productivity to reduce the IT backlog and meet SLAs
- ***Built-in optimizations*** to maximize Hadoop’s vertical scalability; this helps you meet performance requirements and reduce costs by processing more data in less time with fewer nodes
- ***No code generation*** so that you don’t bring traditional data integration shortcomings into Hadoop, such as longer development cycles due to constant fine-tuning, lower performance per node due to inefficient code, and higher hardware costs due to sub-optimal use of resources
- ***Build a proof of concept*** and ‘stress test’ your approach using Hadoop for ETL and other ETL tools with Hadoop to verify that you can meet SLAs, cost, scalability, and connectivity requirements
- ***A cloud solution*** to easily migrate heavy ETL workloads. Some organizations may want to consider a cloud approach such as Amazon Elastic MapReduce (EMR) to instantly procure Hadoop clusters. However, they will still need the right tools to get productive and deliver results quickly in the cloud.

HOW SYNCSORT CAN HELP

Syncsort Hadoop ETL solutions are designed to help organizations realize their business objectives when using Hadoop for ETL initiatives.

- The only approach that runs natively within MapReduce, delivering faster performance and efficiency per node
- Not a code generator; MapReduce invokes Synscort for Hadoop ETL at runtime which executes on all nodes as an integral part of Hadoop
- A “no coding” approach; complex Java, Pig or HiveQL code is replaced with a powerful, easy-to-use graphical development environment
- The first and only Hadoop ETL-as-a-service solution for Amazon EMR, accelerating productivity while leveraging the massive scalability of the Amazon cloud

Step Two

GET THE RIGHT CONNECTIVITY FOR HADOOP

With your business objectives defined, the next step is to ensure you have all the connectivity you need so that Hadoop doesn't become another silo within your organization. Big Data comes from a big list of data sources and targets, including relational databases, files, CRM systems, social media, etc. And that data needs to move in and out of Hadoop, which isn't trivial, requiring manually writing custom scripts with numerous purpose-specific tools such as Sqoop for relational database tables, Hadoop fs shell for files, and Flume for ingesting logs.

Organizations also need to factor in other sources such as legacy and mainframes. With at least 70% of the world's transactional production applications running on mainframe platforms, the opportunity to tap into that data for deeper analytics at a lower cost with Hadoop is important. Yet Hadoop offers no native support for mainframes, which makes it a cumbersome, manual process.

Shortening load times into the Hadoop Distributed File System (HDFS) can be critical for organizations dealing with billions of records every day.

As applications and business needs grow, the need to reduce load times becomes even more important. Of course, streamlined connectivity to the EDW is essential. However, connectivity approaches that involve lots of disparate tools and hand coding mean every time something changes, IT teams need to spend a significant time and effort, which hinders time-to-insight. Therefore, being able to efficiently connect to all sources with a single tool and without coding is key. In addition, capabilities to pre-process data to sort, cleanse, filter, and compress it will further enhance performance and save on storage space.

BEST PRACTICES

Extracting value from Big Data requires extensive data connectivity. The easier and faster you make data connectivity, the more value you'll derive from your data and from your Hadoop investment.

- Select a tool with a *wide variety of connectors,* including relational, cloud, files, and mainframe sources to simplify importing and exporting data to and from Hadoop
- Identify an approach that lets you *pre-process data* for greatest efficiency
- Help to *future-proof* your approach with partners committed to open APIs and open source

HOW SYNCSORT CAN HELP

Syncsort Hadoop ETL solutions include connectivity capabilities critical to successful Hadoop ETL deployments.

- One tool to connect all data sources and targets including relational databases, appliances, files, JSON, XML, cloud, and even mainframe
- Connects Hadoop to all your data without coding
- Pre-processes data prior to loading it into Hadoop for performance and storage savings
- Unique capabilities to read, translate, and distribute mainframe data with Hadoop
- Data Connector APIs providing unlimited opportunities to tap new data sources and targets as needs evolve

Step Three

IDENTIFY THE TOP 20% OF ETL/ELT WORKLOADS

At this point you're ready to begin, but how do you prioritize and focus your efforts for maximum impact? For years, many organizations have struggled with cost and processing limitations of using their EDW for data integration. Once considered best practices, staging areas have become the "dirty secret" of every data warehouse environment – one that consumes the lion's share of time, money, and effort. That's why many Hadoop implementations start with ETL initiatives.

With cheap storage, high reliability, and massive scalability, Hadoop can in fact become the ideal staging area for all your data. Handling transformations and batch processing in Hadoop can easily overcome the critical shortcomings of conventional data integration. But to prove the value Hadoop can bring and build momentum and executive-level support for Hadoop, early success rests in identifying which data transformations to target first.

Usually the top 20% of ETL/ELT workloads can consume up to 80% of the processing capacity, draining significant costs and IT effort due to hardware, tuning, and maintenance.

Targeting these transformations first will achieve the best ROI with the fastest time to value. Focusing on this 20% optimizes results of Hadoop ETL efforts. Operational savings and deferred database costs can then be used to fund more strategic initiatives.

BEST PRACTICES

When identifying the top 20% of transformations to target for migration to Hadoop, look for jobs with one or more of these challenges:

- ***Relatively high elapsed processing times***
- ***Very complex scripts,*** including change-data-capture (CDC), slowly changing dimensions, raking functions, volatile tables, multiple merge, joins, cursors and unions
- ***Files and semi-structured data*** such as web logs and clickstream analysis
- ***High impact on resource utilization,*** including CPU, memory, and storage
- ***Unstable or error-prone code***

HOW SYNCSORT CAN HELP

Using Syncsort solutions to jumpstart ETL through Hadoop enables organizations to:

- Analyze and understand complex SQL scripts via an intelligent, self-documenting flow chart
- Migrate thousands of lines of code to a few graphical Syncsort Hadoop ETL jobs
- Quickly identify and troubleshoot data errors through comprehensive logging and debugging capabilities

Step Four

RE-CREATE EQUIVALENT TRANSFORMATIONS IN MAPREDUCE

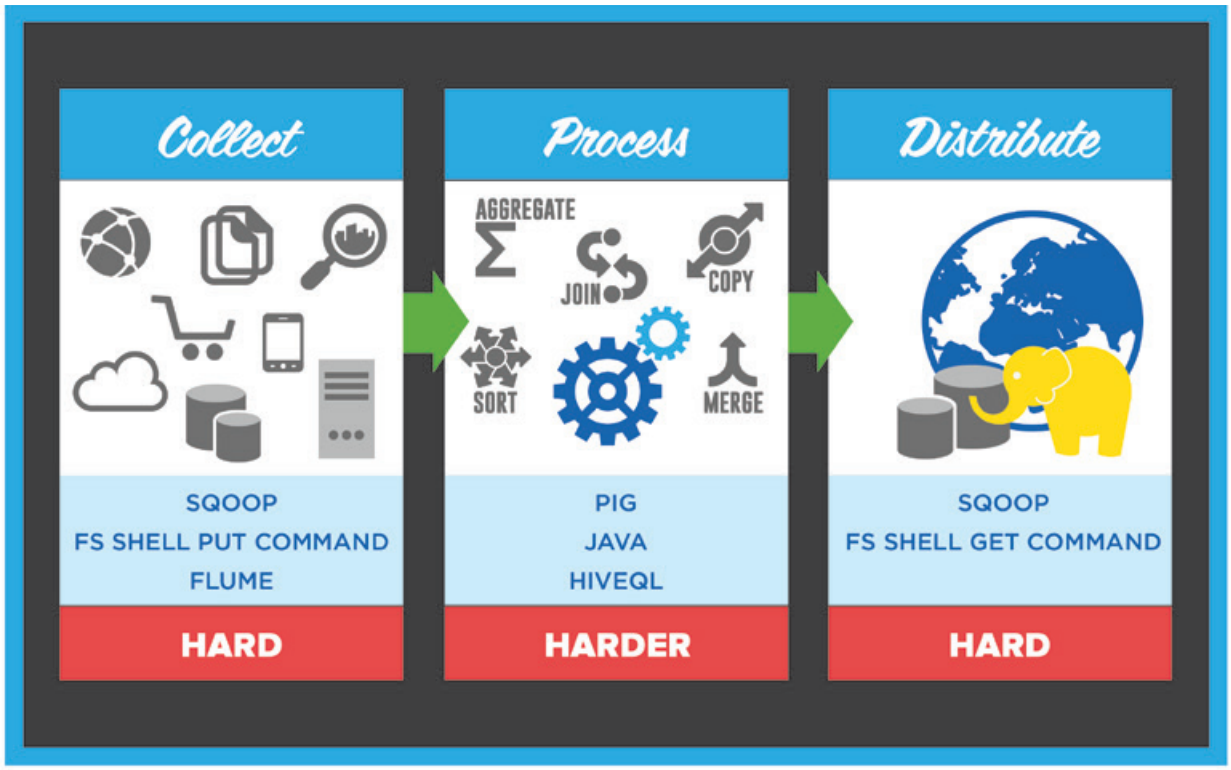
Now that you've identified which ETL workloads to start with, your next step is to understand all the data transformations happening in the EDW and then how to replicate them in Hadoop. Proprietary SQL variants like Oracle PL/SQL with limited metadata and documentation make this particularly challenging.

Once you do navigate the complexities of SQL-based ETL jobs, a rich ecosystem of Hadoop utilities are available to create ETL jobs, but they are all separately evolving projects and require specific, new skills.

Developers need to be well-versed in Java, HiveQL, and Pig for developing meaningful MapReduce ETL jobs. Not only do these tools require hand coding, which reduces productivity, but in most cases you will also need a deep understanding of Hadoop and MapReduce, especially when there's the need for user-defined functions (UDF).

Moreover, some scripts can result in additional overhead and even when written by expert developers will typically require numerous iterations to achieve optimal performance. For example, HiveQL statements need to be translated in MapReduce jobs before being submitted into Hadoop, adding overhead.

ETL IN HADOOP TODAY



Data transformations can quickly become complex with Hadoop. The gaps between Hadoop and enterprise ETL – tasks like sort, joins, aggregations, and connectivity – require complex manual coding, in effect reverting to the inefficiencies and high costs of traditional data integration.

Building more sophisticated data flows such as a CDC process, widely used in ETL today, is even more difficult in Hadoop. Data sets are typically much larger and distributed across data nodes in HDFS – records need to be co-located to identify changes; and then a great deal of hand coding and tuning (easily hundreds of lines of manual code) is required to achieve acceptable performance.

BEST PRACTICES

Use short-cuts to begin the process of offloading your EDW:

- *Start by analyzing, understanding and documenting* complex transformations and processes occurring in the EDW
- *Leverage tools with point-and-click interfaces* to quickly develop common ETL use cases and start migrating those first
- *Avoid tools that require manual coding*
- *Ensure you can leverage existing programming skills* with tools that offer pre-built data integration functions and a graphical user interface

HOW SYNCSORT CAN HELP

Syncsort Hadoop ETL solutions are mature data integration tools that make it easy to create and maintain Hadoop ETL jobs via:

- The ability to develop and test locally in a Windows-based graphical user interface, then deploy in Hadoop
- Comprehensive built-in transformations with MapReduce jobs with Mappers and Reducers out-of-the-box
- A library of Use Case Accelerators to overcome a steep learning curve and quickly develop common ETL tasks such as CDC, aggregations, joins, and more in Hadoop
- Built-in metadata capabilities for increased reusability, impact analysis, and data lineage

Step Five

MAKE YOUR HADOOP ETL ENVIRONMENT ENTERPRISE-READY

As you shift more ETL processes to Hadoop, you now need to lay a solid foundation for ongoing success. There's no quicker way to thwart your transition to Hadoop than by failing to meet deadlines and missing performance SLAs for ETL processing. In addition, the organization expects the same level of functionality and services provided by their conventional enterprise data integration tools, only faster and less costly. Hadoop is lowering the cost structure of processing data at scale. However, deploying Hadoop at the enterprise level is not free and significant hardware and IT productivity costs can damage ROI. Although Hadoop leverages commodity hardware, when dealing with large numbers of nodes, hardware costs add up. Programming resources – e.g., HiveQL, Pig, Java, MapReduce – can also prove expensive and compromise productivity.

Many data integration tools work peripherally to Hadoop – generating Hive, Pig or Java – adding a layer of overhead that hurts performance. ETL solutions that are tightly integrated with Hadoop and avoid code generation are easier to deploy and maintain with no performance impact or hurdles down the road.

One of the challenges with conducting data transformations in an EDW is a lack of metadata, which makes impact analysis, job tracking and reusability impossible. BTEQ scripts which routinely contain thousands of lines of code have to be hand coded again for each job and maintained manually. Ensuring metadata capabilities as part of Hadoop ETL to simplify management and reusability is essential when meeting SLAs.

Information is one of the most valuable assets of any organization and with Big Data comes even bigger responsibility. Therefore, the ability to maintain enterprise-level data security in Hadoop is also critical, yet capabilities to secure data integration processes in Hadoop are limited. While some ETL tools offer graphical user interfaces and connectivity, they provide their own security models, which can be difficult to integrate and synchronize with your own enterprise standard. Support for tight security requirements using existing security infrastructure is essential.

BEST PRACTICES

Ensuring SLAs as you offload your EDW by using Hadoop for ETL will help pave the way for future Hadoop initiatives. To do this:

- Understand how different solutions specifically *interact with Hadoop* and the type and *amount of code* they generate
- Identify an approach that complements the *benefits of open source* to deliver savings and efficiencies
- Consider a tool with *native Hadoop integration* to meet performance SLAs and avoid unnecessary overhead
- Seek solutions that offer a *metadata repository* to enable re-use of developments and data lineage tracking
- Make sure security isn't compromised. Any viable approach must leverage existing infrastructure to *control and secure all your data*
- Look for tools that offer *scalable approaches* to deploy, monitor, and administer your Hadoop ETL environment

HOW SYNCSORT CAN HELP

Syncsort Hadoop ETL solutions deliver an enterprise-ready approach to Hadoop ETL.

- Run natively within Hadoop; the runtime engine executes on all nodes as an integral part of the Hadoop framework
- Tightly integrate with all major Hadoop distributions, including Apache, Cloudera, Hortonworks, MapR, PivotalHD and even Amazon EMR
- Seamlessly integrate with Cloudera Manager for one-click deployment and upgrade of Syncsort Hadoop ETL solutions across the entire Hadoop cluster
- Provide full integration with Hadoop Jobtracker for easier monitoring of MapReduce ETL jobs
- Plug into existing Hadoop clusters to seamlessly optimize existing HiveQL and MapReduce jobs for even great performance and more efficient use of Hadoop clusters
- Automatically self-optimize based on resources and tasks to deliver faster, sustainable performance and efficiency per node with a Smart ETL Optimizer
- Keep data secure with market-leading support for common authentication protocols such as LDAP and Kerberos

Conclusion

Organizations are shifting ETL from the EDW to Hadoop in order to reduce costs and free up database capacity for faster analytics and end-user queries. But Hadoop is not a complete ETL solution – its primary intent is as an operating system for Big Data. For optimal results it must be coupled with an enterprise ETL solution.

Whether you choose Syncsort Hadoop ETL solutions, or another approach, these five steps can help you offload your EDW to Hadoop, overcome some of the biggest challenges, and arm you with some best practices to accelerate data integration efforts and ensure success for future Hadoop initiatives.

Step 1:

Understand and define business objectives. Whether you're aiming to reduce costs, leverage more data, or meet SLAs, it's critical to understand your primary objectives to ensure that you meet them as you shift heavy ETL workloads out of the EDW and into Hadoop. An approach that leverages existing developer skill sets, provides enterprise ETL functionality, and runs natively in Hadoop to optimize ETL performance helps you to reach your goals.

Step 2:

Get the right connectivity for Hadoop. Make sure Hadoop doesn't become another silo within your organization. To extract value from Big Data you need extensive data connectivity and Hadoop requires significant manual effort to make this happen. With a tool that provides connectivity to all data sources and targets without coding, you gain more value from your data and from your Hadoop investment, faster.

Step 3:

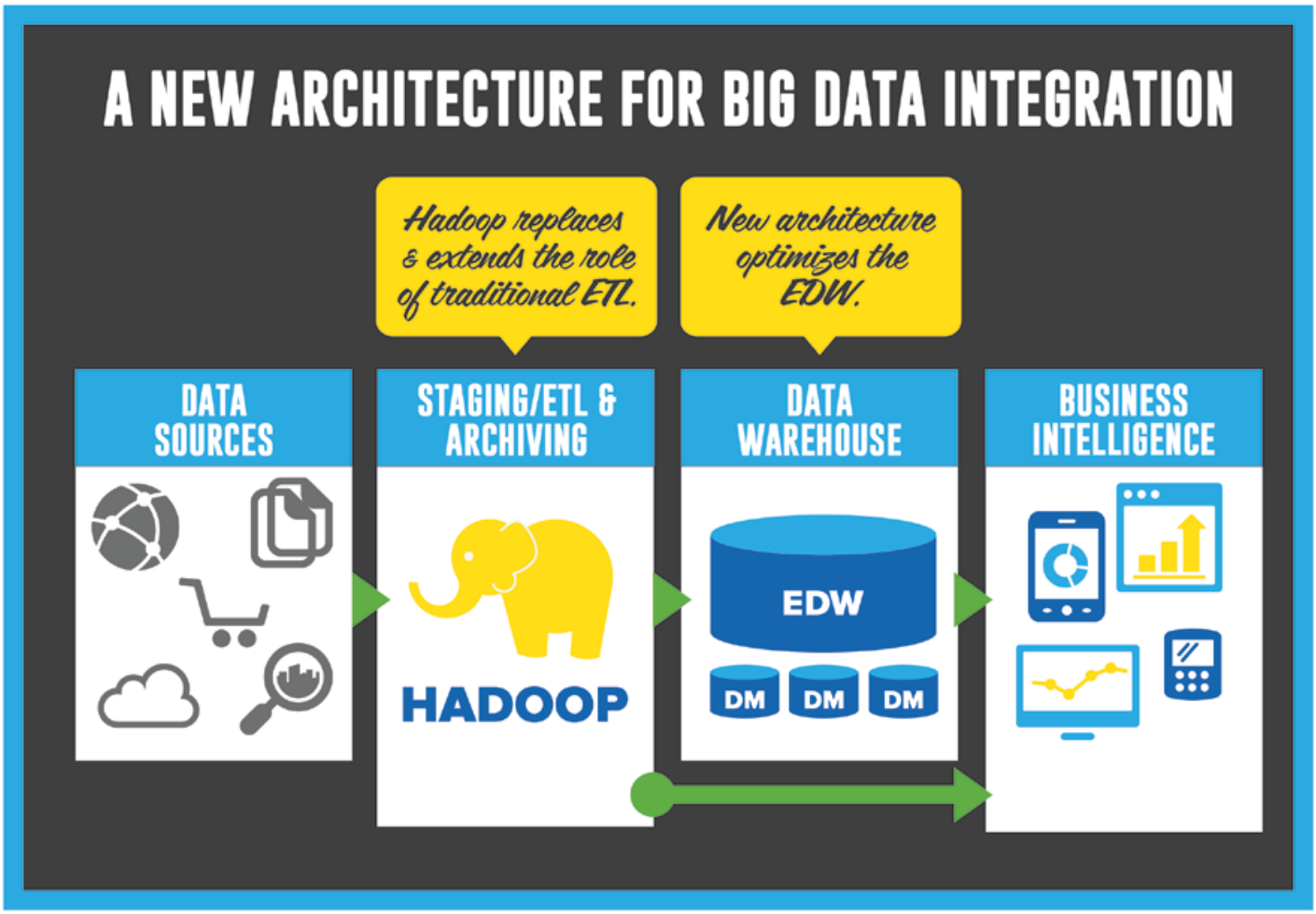
Identify the top 20% of ETL/ELT workloads. Usually the top 20% of data transformations can consume up to 80% of the processing capacity, draining significant costs and IT effort due to hardware, tuning, and maintenance. Start with these workloads and you'll achieve the best ROI with the fastest time to value, optimizing results of Hadoop ETL efforts.

Step 4:

Re-create equivalent transformations in MapReduce. Data transformations can quickly become complex in Hadoop, requiring expertise in Java, Pig, and HiveQL, as well as a deep understanding of Hadoop and MapReduce, and significant manual coding and tuning. A GUI-based enterprise ETL approach with transformations, common ETL tasks, and metadata capabilities all built-in make it easy to create and maintain Hadoop ETL jobs.

Step 5:

Make your Hadoop ETL environment enterprise-ready. Failure to meet SLAs for ETL processing is a sure way to torpedo your Hadoop initiative. Capabilities that facilitate large-scale deployments, monitoring and administration, and data security in Hadoop deliver an enterprise-ready approach that paves the way for future initiatives. These capabilities include native Hadoop integration, built-in performance optimization, and the ability to leverage existing security infrastructure.



ABOUT US

Syncsort provides data-intensive organizations across the big data continuum with a smarter way to collect and process the ever-expanding data avalanche. With thousands of deployments across all major platforms, including mainframe, Syncsort helps customers around the world to overcome the architectural limits of today's ETL and Hadoop environments, empowering their organizations to drive better business outcomes in less time, with less resources and lower TCO. For more information visit www.syncsort.com.

syncsort

LIKE THIS? SHARE IT!