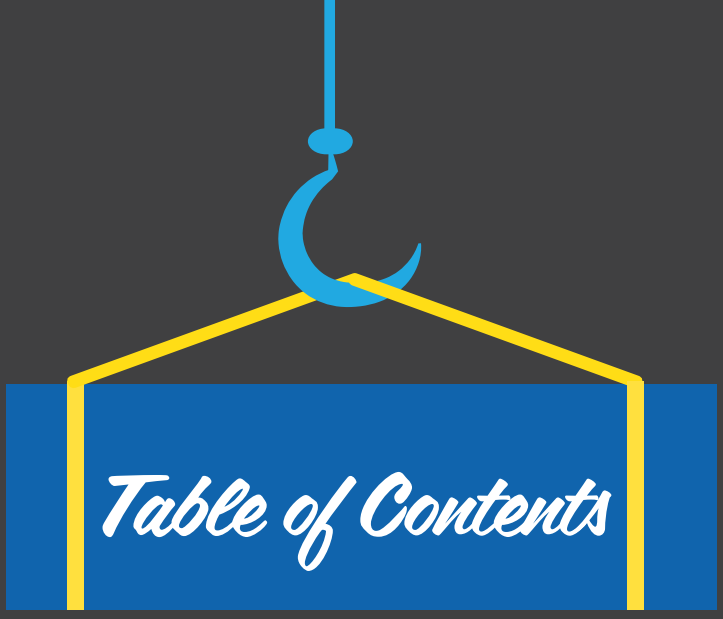
NEW HADOOP GUIDE!

A 1-2-3 APPROACH TO OFFLOADING TERADATA

With Hadoop

A Practical Guide to Freeing up Valuable Teradata Capacity & Saving Costs with Hadoop

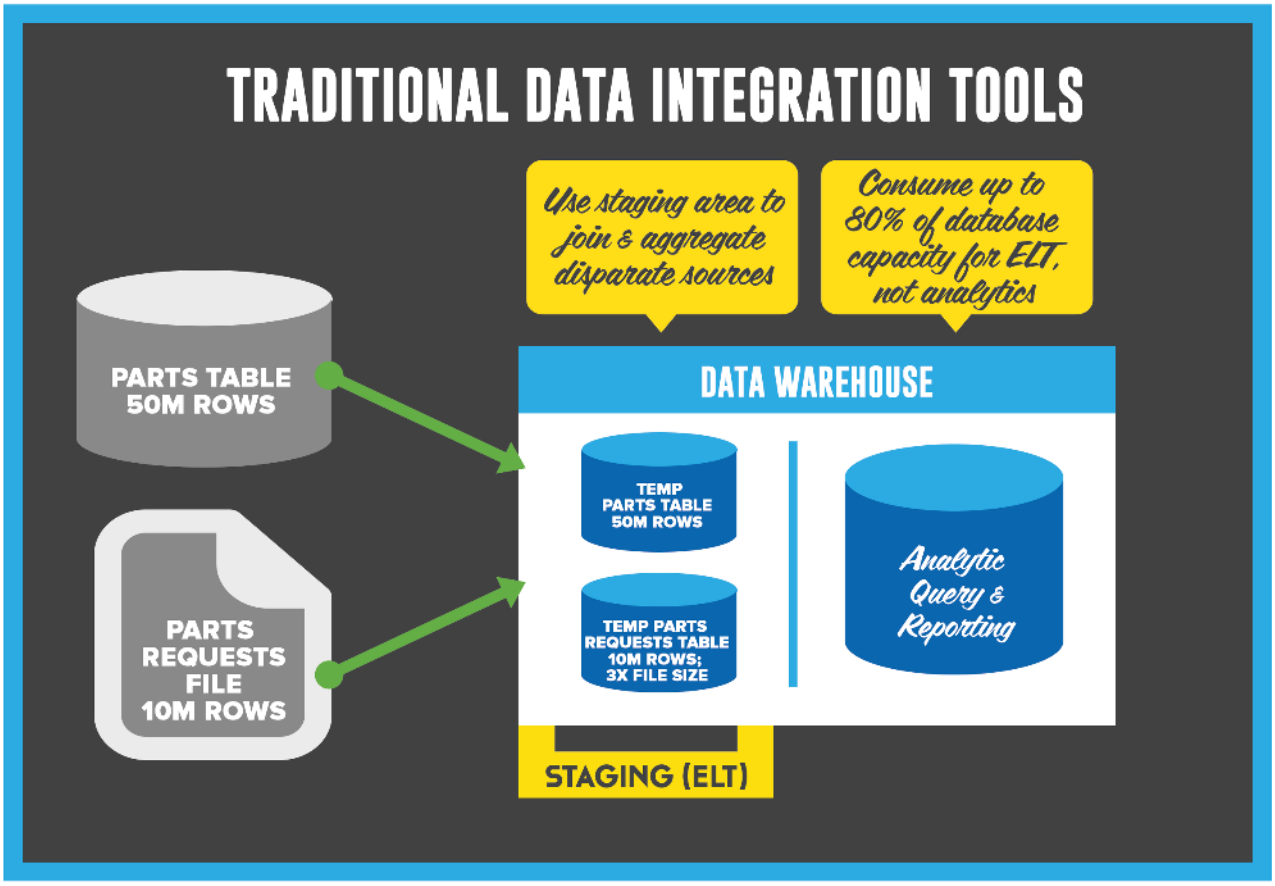
syncsort



THE PERVASIVE IMPACT OF ELT

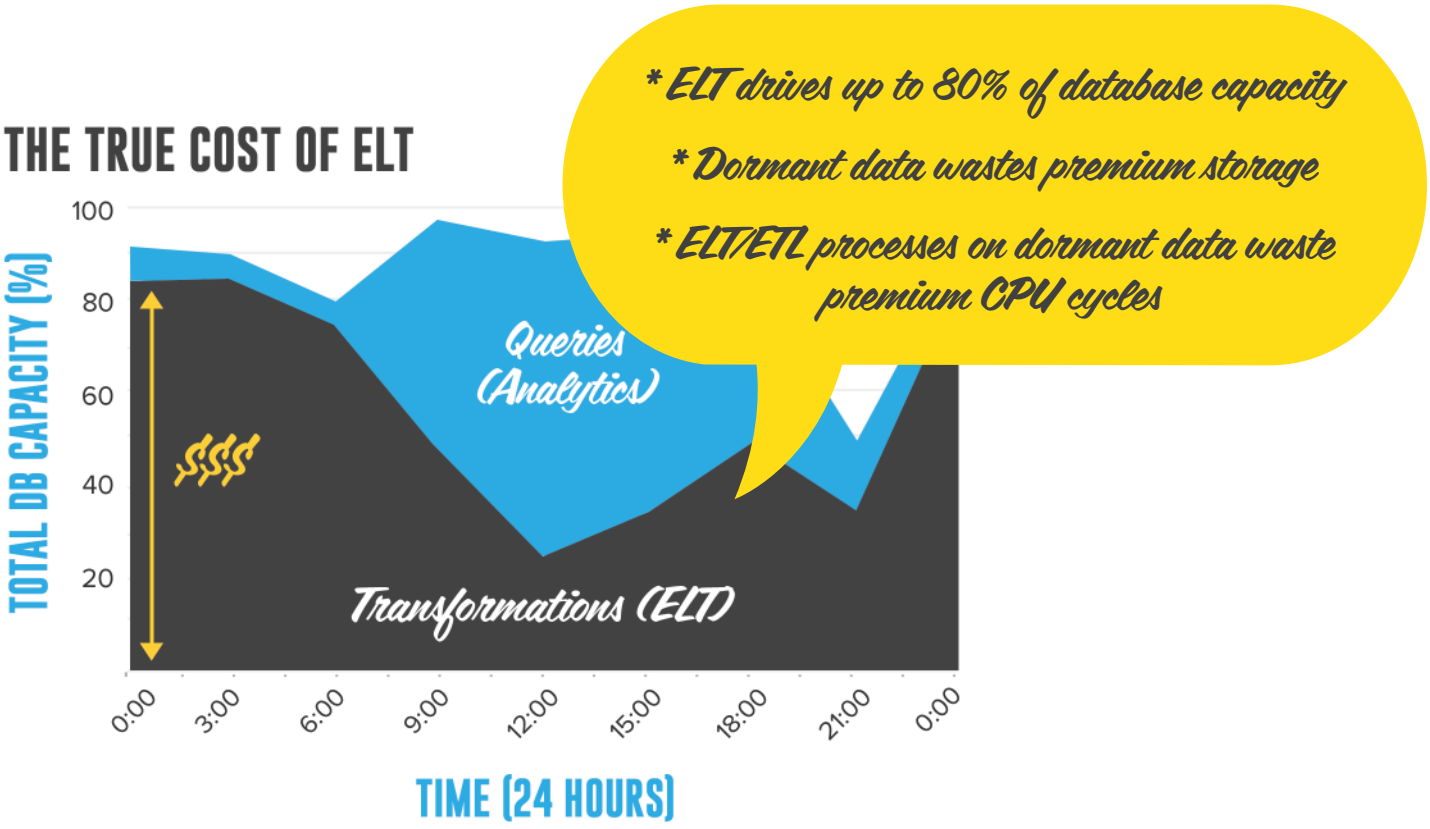
At an event in late 2013 to unveil grants to support Big Data programs at prominent universities, John Holdren, director of the White House Office of Science and Technology Policy stated, "Big data is now a super-big deal." The ability to make sense of Big Data is so prized that a growing list of educational institutions are now offering degrees in the discipline. With the universe of data measuring more than 3 trillion gigabytes and expected to double every two years, the focus on big data is understandable. 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 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, conventional data integration approaches forced IT to push the transformations down to enterprise data warehouses, like Teradata, creating a shift from ETL to ELT. But this proved to be a costly and inefficient approach.



Enterprises are asking for fresher data – from daily, to hourly, to real-time – as well as access to historical data from more sources and for longer periods of time, and they need it faster and cheaper. ELT simply can't satisfy these requirements. According to Gartner, nearly 70% of all data warehouses are performance and capacity constrained. Longer batch windows – from 4 hours, to 8, and even 12 hours – are quickly becoming the norm, competing with the actual intended use of Teradata (reporting and analytics) and hampering response times.

With costs and data volumes growing, data retention capacity measures 3 months at best, making it impossible to capture the insights longer data histories can provide. A pressing requirement of Big Data, connecting new data sources – structured, unstructured or semi-structured – to the breadth of analysis capabilities can take months. Simply adding a single new metric or column is burdensome. Organizations quickly find themselves in an unsustainable position with costs outpacing data volume. It's no surprise that many cite total cost of ownership as their #1 challenge with their data integration tools.

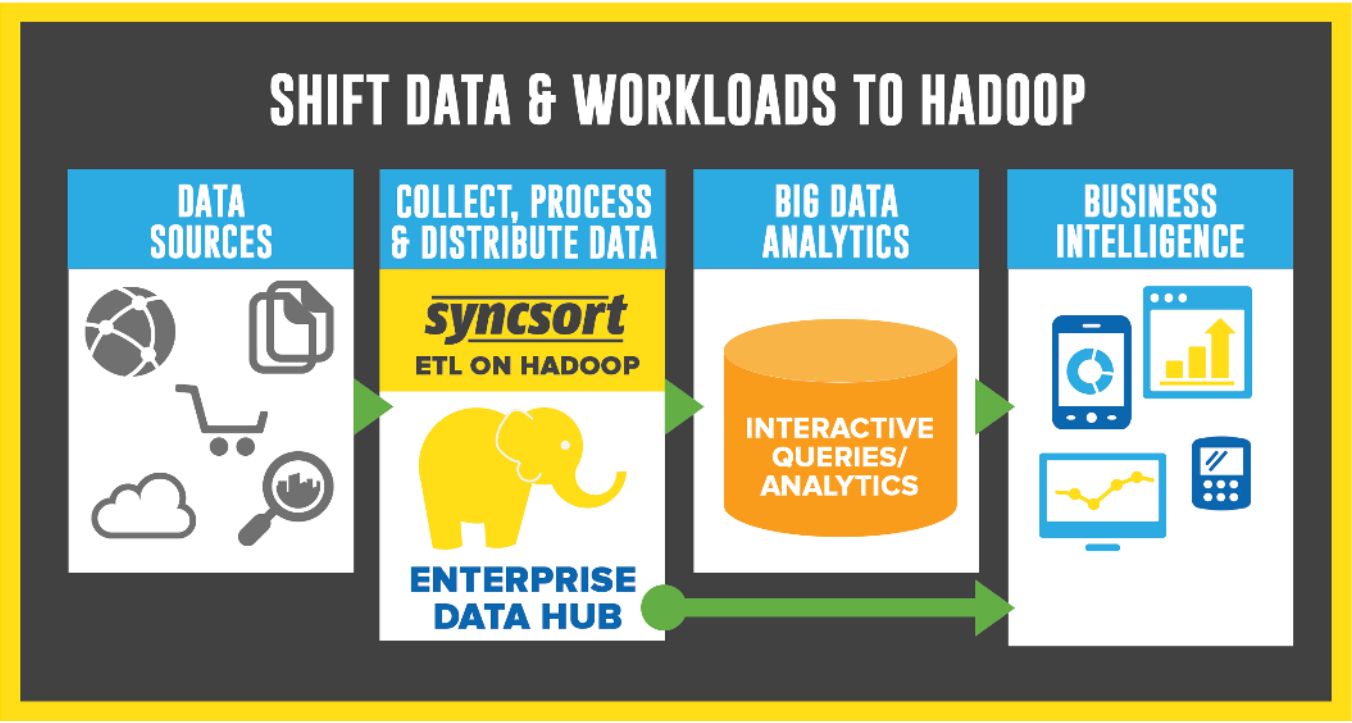


With no end in sight to the digital explosion, organizations are more recently looking at Hadoop as the perfect staging area to collect and transform data from disparate sources, before selectively loading key datasets into Teradata for valuable analytics, operational intelligence, and end-user queries. By shifting ETL to Hadoop – offloading heavy transformations from Teradata – organizations are finding they can dramatically reduce costs, incorporate new data faster, and free up Teradata capacity for faster analytics and end-user response times.

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 and free up storage and processing power from premium platforms like Teradata. Estimates from multiple sources indicate managing data in Hadoop can range from $500 to $2,000 per terabyte of data, compared to $20,000 to $100,000 per terabyte for high-end data warehouses. Meanwhile, even Teradata recognizes that ELT workloads consume a significant amount of data warehouse resources.[1] Recapturing some of that capacity provides a huge opportunity to defer the costs of a Teradata upgrade while making the most of newfound data warehouse capacity. With ELT driving up to 80% of Teradata resource utilization, the upside can be significant.

Shifting ETL to Hadoop not only saves costs but helps IT departments be more agile and work more efficiently, reallocating Teradata resources – hardware as well as highly-skilled Teradata developers and business analysts – toward high-value activities such as sophisticated analytics and interactive queries. High-performance and high-scalability are key reasons why organizations select Teradata as their data warehouse. Unfortunately, these advantages quickly get lost in a fog of ELT workloads. Using Hadoop to allocate workloads appropriately enables organizations to realize the advantages of reliability, accuracy, consistency, and low latency that Teradata delivers.



1 http://blogs.barrons.com/techtraderdaily/2013/10/31/teradata-slips-q3-q4-remarks-prompt-great-hadoop-debate/

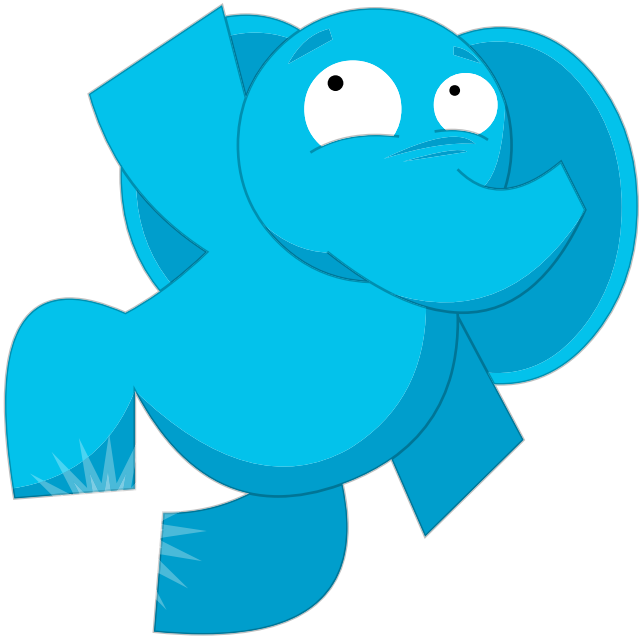
But Hadoop is not a complete ETL solution. While Hadoop offers powerful utilities and virtually unlimited horizontal scalability, it does not provide the complete set of functionality users need for enterprise ETL. In most cases, these gaps must be filled through complex manual coding, slowing Hadoop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

However, there is a way to combine the benefits of purpose-built, high-performance ETL with Hadoop, optimizing your Teradata data warehouse while gaining all the benefits of a complete ETL solution.

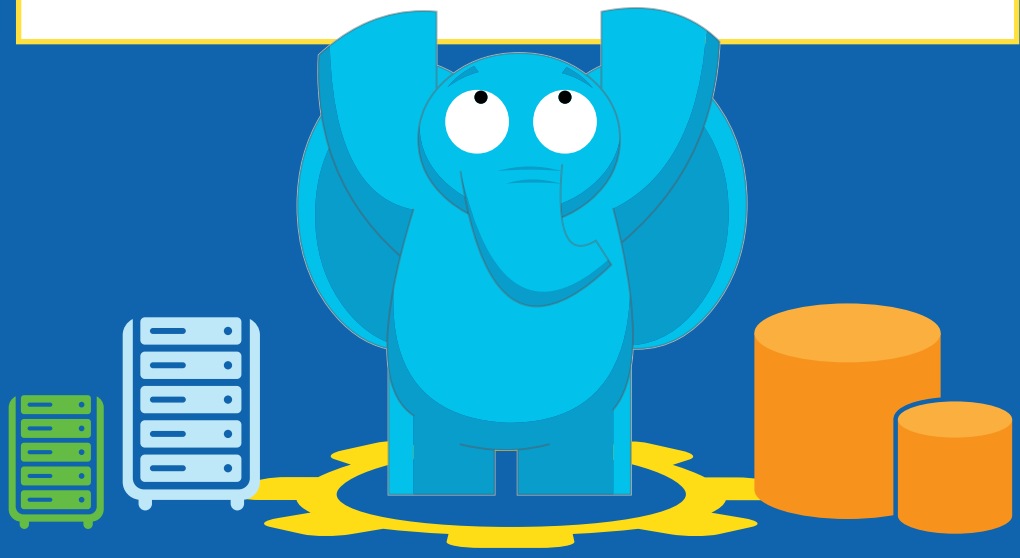
This guide will provide a three-phased approach to help you overcome some of the biggest challenges of creating an enterprise data hub on Hadoop, and arm you with some best practices to accelerate data integration efforts. By following these steps you'll be able to:

- Gain fresher data, from right time to real time
- Keep all data as long as you want, up to months, years and beyond
- Optimize your data warehouse for faster queries and faster analysis
- Blend new data fast, from structured to unstructured, mainframe to the Cloud
- Free up your budget, from $100K per terabyte to $2K per terabyte

The cost savings alone will allow you to justify the investment in Hadoop and help you to gradually build up your organization's Hadoop skills. And the additional capacity will enable you to focus more resources on interactive user queries, speed-of-thought analytics, and business intelligence so you can capitalize on the key strengths of your Teradata platform.



BUILDING THE ENTERPRISE DATA HUB



Most organizations have spent considerable time and money building their existing data warehouse architecture. Once considered best practices, these architectures called for large and expensive data integration staging areas to blend data from disparate sources. But increasing demands for information have proven to be too much for this approach to handle on a number of fronts:

- **COSTS** – Relying on Teradata 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 $500K per year on additional Teradata capacity...just to keep the lights on!

- **DATA VOLUME AND VARIETY** – 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 a combination of Teradata 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 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 to integrating data.

Realizing the critical value of data to sustain competitiveness, Hadoop has become the framework of choice to store all the data available to the organization. Using Hadoop as the central repository for all kinds of data, organizations can accelerate time-to-insight and reduce the overall costs of collecting, processing and distributing data. 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 and approach, 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.

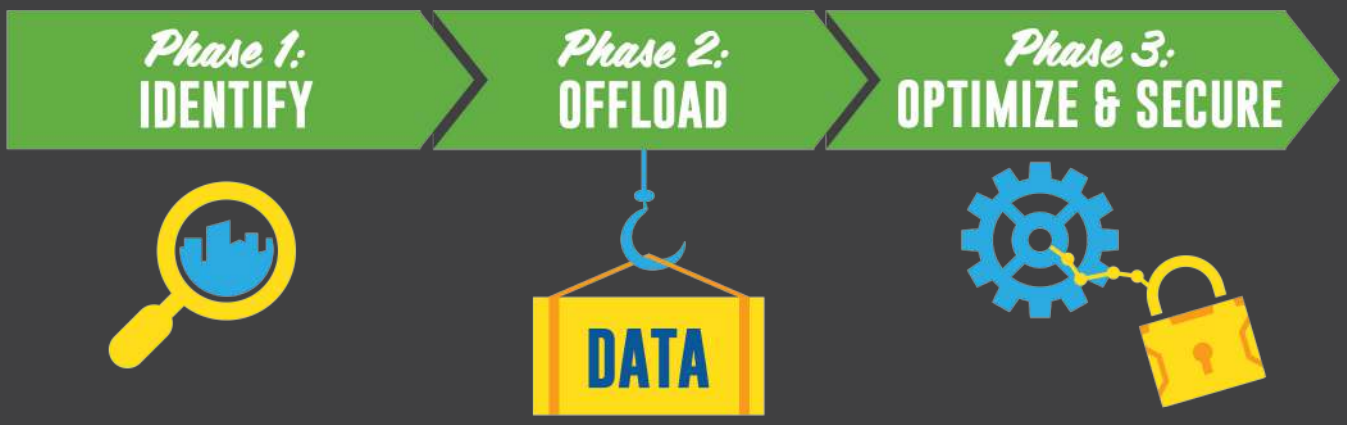
Some of the main challenges of using Hadoop as your main ETL tool include:

- Finding and paying for skilled Hadoop programmers with knowledge of Java, Pig, Hive, and Sqoop
- Jeopardizing gains in productivity as Hadoop lacks traditional "enterprise ETL" functionality – point and click interfaces, metadata, re-usability, and connectivity – forcing a return to complex coding and the associated development delays, maintenance, and reuse headaches
- Hindering performance by supplementing Hadoop with tools that introduce additional overhead and can't leverage and optimize the Hadoop architecture

The following three-phased approach can help you overcome the challenges of offloading data and ELT workloads from Teradata to Hadoop:

- **PHASE I:** Identify infrequently used – "cold" and "warm" – data. Pinpoint heavy workloads to prioritize for transformation. In most cases 20% of data transformations consume up to 80% of resources.
- **PHASE II:** Offload expensive data and workloads to Hadoop quickly and securely with a single tool. Easily replicate existing workloads without coding.
- **PHASE III:** Optimize & secure the new environment. Optimize performance for maximum throughput per node. Manage and secure all of your data with enterprise-class tools.

THE SYNCSORT OFFLOAD FRAMEWORK



LET'S EXPLORE HOW EACH OF THESE PHASES WILL HELP ENSURE THAT YOU LAY A STRONG FOUNDATION FOR AN ENTERPRISE DATA HUB, ENABLING YOU TO ACHIEVE YOUR BUSINESS OBJECTIVES

PHASE 1: IDENTIFY

You know data and workloads exist in your Teradata data warehouse that could be offloaded, but how do you identify them?

Organizations are increasingly struggling with cost and processing limitations of using their Teradata data warehouse for ELT. 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 ELT offload initiatives. With cheap storage, high reliability, and massive scalability, Hadoop can in fact become the ideal staging area for all your data, a solid first step towards building your enterprise data hub. Handling transformations and batch processing in Hadoop can easily overcome the critical shortcomings of conventional data integration. But to prove the value of Hadoop and build momentum and executive-level support, early success rests in identifying which data transformations to target first.

Usually the top 20% of ELT workloads can consume up to 80% of resources, draining significant costs and IT effort due to hardware, tuning, and maintenance. Dormant data makes the problem even worse. Some data warehouses can contain significant volumes of rarely used data, wasting premium storage capacity and, even worse, consuming a tremendous amount of CPU on ELT workloads for data that the business doesn't actively use.

Targeting these types of data and transformations first will achieve the best ROI with the fastest time to value and optimize 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:

- Dormant data that isn't actively used but must be maintained
- 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 click-stream analysis
- High impact on resource utilization, including CPU, memory, and storage
- Unstable or error-prone code

In addition:

- Look for utilities that automatically identify dormant data for active archive on Hadoop to accelerate your ability to recover premium Teradata storage
- Leverage tools that provide integrated analysis and documentation of complex transformations and processes occurring in the data warehouse

IDENTIFY

WHAT SYNCSORT DOES FOR YOU:

Integrated analysis of EDW data & workloads to identify:

- Expensive transformations
- Unused data
- Rarely used historical data
- Expensive user activity

PHASE II: OFFLOAD

Do you have all the tools and skills necessary to access and move your data and processing?

An enterprise data hub cannot be a silo; connectivity to Hadoop is critical. Big Data comes from a big list of data sources and targets, not just your Teradata data warehouse but also relational databases, files, CRM systems, web logs, social media, mainframes and legacy systems, new file formats such as JSON, etc. And that data needs to move in and out of Hadoop, which isn't trivial, requiring manually written custom scripts with numerous purpose-specific tools such as Sqoop for relational database tables, Hadoop fs shell for files, and Flume for ingesting logs.

Additionally, Hadoop offers no native support for mainframes, requiring a cumbersome, manual process. A majority of organizations in data-intensive industries like financial services, telecommunications, and retail have historically relied on mainframes for transactional applications and Teradata for analytics. It is critical to have an easy and secure way to bring mainframe data into Hadoop for archiving and processing, which can then feed relevant datasets to Teradata. Connectivity approaches that involve lots of disparate tools and hand coding mean every time something changes, highly-skilled Teradata analysts need to spend significant time and effort, which hinders time-to-insight.

Replicating transformations created using BTEQ scripts into Hadoop is another challenge. 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 into MapReduce jobs before being submitted into Hadoop, adding overhead.

Data transformations can quickly become complex with Hadoop. The gaps between Hadoop and enterprise ETL – connectivity and tasks like sort, joins and aggregations – 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:

Extracting value from Big Data requires extensive data connectivity and efficient processing. The easier and faster you make it to access, move, and process data, 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
- Adopt an approach that lets you pre-process data to sort, cleanse, filter, and compress it before loading into Hadoop for greatest efficiency
- Leverage tools with point-and-click interfaces to quickly develop common ETL use cases and start migrating those first
- Avoid tools that generate code or require manual coding
- Ensure you can leverage existing programming resources with tools that offer pre-built data integration functions and a graphical user interface

OFFLOAD

WHAT SYNCSORT DOES FOR YOU:

Access virtually any data, anywhere with one tool:

- Extract data & load into HDFS natively from Hadoop: RDBMS, JSON, XML, files, mainframe, cloud, & more
- Cleanse, validate, partition, & compress

Translate ELT workloads to optimized MapReduce processes, without coding:

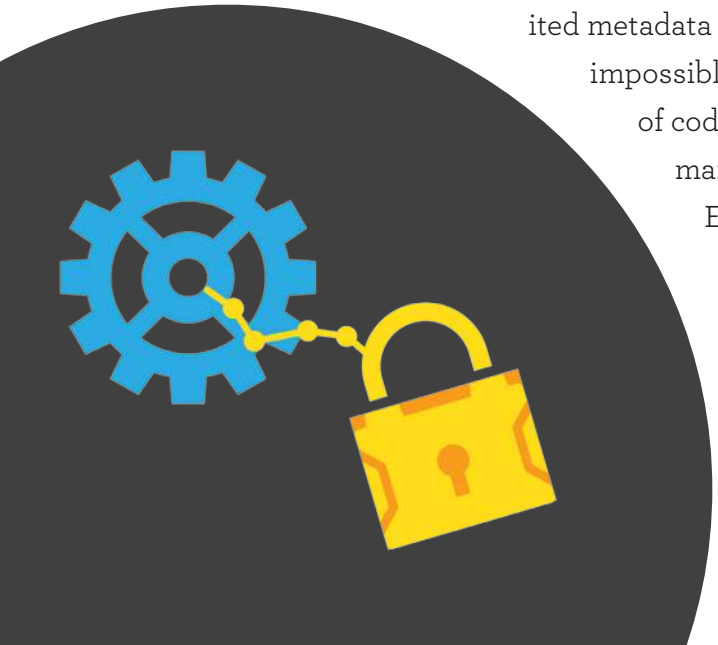
- Develop MapReduce processes graphically
- Develop & test locally in Windows
- Fast-track development with reusable templates
- Create once, re-use many times!

PHASE III: OPTIMIZE & SECURE

How can you optimize your processing inside of Hadoop and ensure you are laying a secure foundation for an enterprise data hub?

As you shift more ELT processes to Hadoop, you now need to make sure you have the tools and processes in place to manage, secure, and operationalize your enterprise data hub for ongoing success. There's no quicker way to thwart your transition to Hadoop than by failing to meet deadlines, missing performance SLAs, or raising security concerns. In addition, the organization expects the same level of functionality and services provided when transformations were done in Teradata, 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. Some of these tools generate code that can be sub-optimal, requiring organizations to spend weeks if not months trying to optimize transformations. Typically this requires additional steps in the graphical user interface (GUI) or finally resorting to hand coding user-defined functions in Java, or other languages. ETL solutions that are tightly integrated with Hadoop, have a built-in optimizer, 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 Teradata is limited metadata which makes impact analysis, job tracking, and re-usability 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 re-usability 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 GUIs 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.

As organizations shift ELT processes to Hadoop, it is important to make sure the Hadoop ETL environment is enterprise-ready. Capabilities that facilitate large-scale deployments, monitoring, and administration will be key to overcome resistance to change and accelerate adoption.

BEST PRACTICES:

Securing your Hadoop environment and ensuring SLAs as you offload Teradata 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
- Select tools with no-hassle support for common authentication protocols such as LDAP and Kerberos to load and extract data to/from Hadoop as well as executing jobs
- Look for tools that offer scalable approaches to deploy, monitor, and administer your Hadoop ETL environment

OPTIMIZE & SECURE WHAT SYNCSORT DOES FOR YOU:

Optimize:

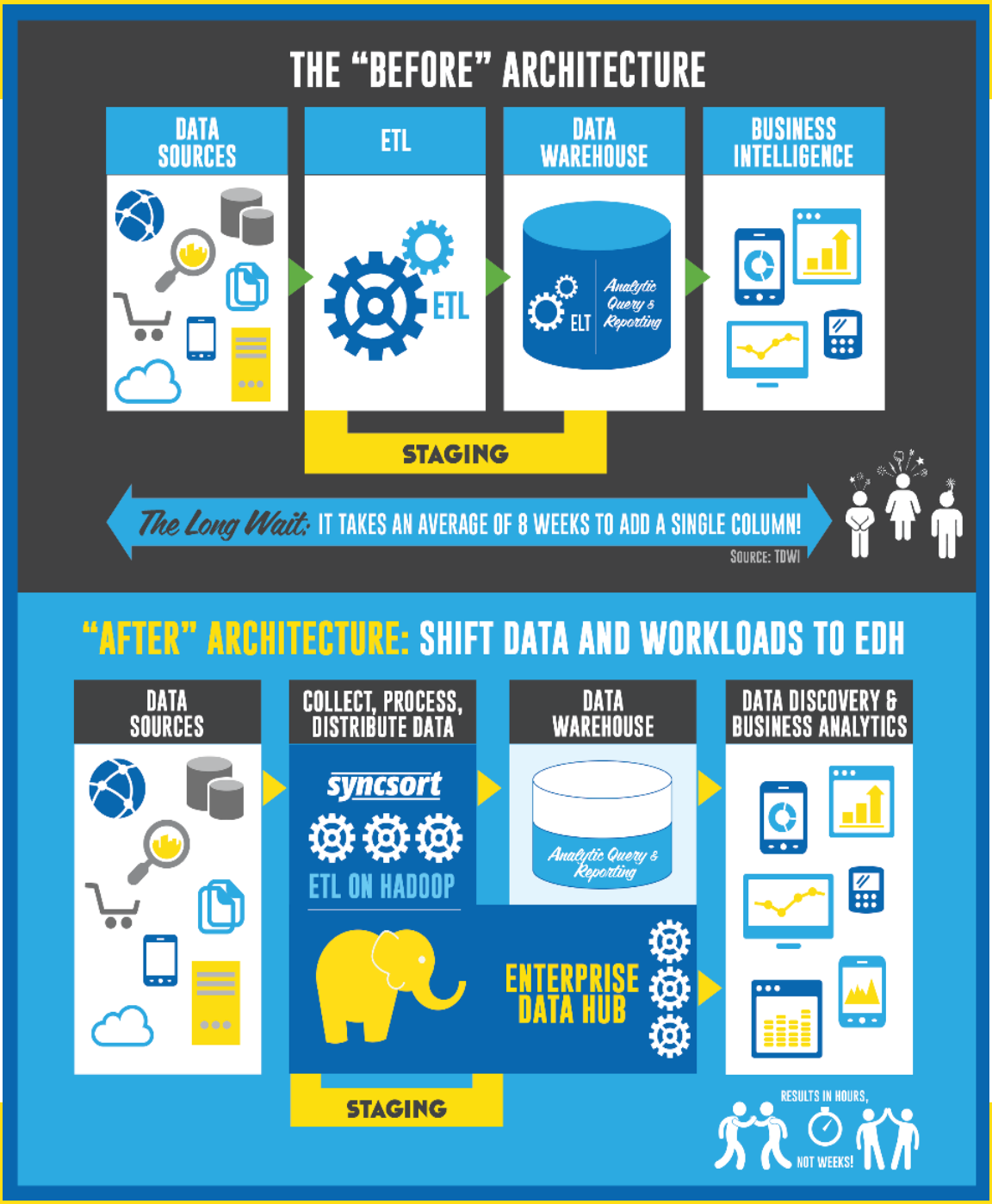
- Get more throughput per node
- Accelerate existing MapReduce, Hive, & Pig processes

Secure:

- Leading support for Kerberos & LDAP
- Secure data loads & extracts
- Secure job execution

SHIFTING ELT WORKLOADS TO HADOOP

Organizations are shifting heavy ELT workloads from Teradata 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. Therefore, offloading data and workloads from Teradata into Hadoop can be intimidating. How do you know where to begin, and what will deliver the most savings? How can you make sure you have all the tools you need to access and move your data and processing? And once you do that, you'll need to optimize all the processes inside Hadoop and guarantee secure data access.

SYNCSORT'S TERADATA OFFLOAD SOLUTION

Syncsort provides targeted offload solutions, specifically designed to address these challenges. Start with Syncsort's offload solutions to save money while building your enterprise data hub to power next-generation big data analytics.

CHALLENGE: IDENTIFY

How do you determine which processes are suitable for offload as well as how many resources target processes consume?

- Integrated analysis of all enterprise data warehouse data and workloads to identify data and processes suitable for offload
- Understand and document complex BTEQ scripts

CHALLENGE: CONNECTIVITY

Teradata ingests data from multiple and diverse sources. How do you bring all these sources into Hadoop?

- 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

CHALLENGE: SECURITY

New data approaches must be secure. How do you ensure secure data movement and access?

- Keep data secure with market-leading support for common authentication protocols such as LDAP and Kerberos

CHALLENGE: EXPERTISE

How do you take BTEQ expertise and make those skills relevant in Hadoop?

- A "no coding" approach; complex Java, Pig or HiveQL code is replaced with a powerful, easy-to-use graphical development environment
- 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
- The ability to develop and test locally in a Windows-based graphical user interface, then deploy in Hadoop
- The first and only Hadoop ETL-as-a-service solution for Amazon EMR, accelerating productivity while leveraging the massive scalability of the Amazon cloud

CHALLENGE: OPTIMIZATION

How will you manage and maintain the new Hadoop-based ETL processes?

- Built-in metadata capabilities for increased re-usability, impact analysis, and data lineage
- Run natively within Hadoop; the runtime engine executes on all nodes as an integral part of the Hadoop framework
- 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; tightly integrate with all major Hadoop distributions, including Apache, Cloudera, Hortonworks, MapR, PivotalHD and even Amazon EMR
- Plug into existing Hadoop clusters to seamlessly optimize existing HiveQL and MapReduce jobs for even greater 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

Your Teradata data warehouse is a key investment; use it wisely – for fast, interactive user queries, speed-of-thought analysis, and business intelligence. Teradata still plays and will continue to play a vital role for many organizations. But by offloading data and heavy ELT workloads from Teradata onto Hadoop you can free up expensive data warehouse capacity for more value-added activities and cost savings. The three-phased approach outlined here can ensure you are laying the foundation for an enterprise data hub that will help you achieve your business objectives.

- **PHASE I: IDENTIFY** infrequently used – "cold" and "warm" – data. Pinpoint heavy workloads to prioritize for transformation. In most cases 20% of data transformations consume up to 80% of resources.
- **PHASE II: OFFLOAD** expensive data and workloads to Hadoop quickly and securely with a single tool. Easily replicate and optimize existing workloads without coding.
- **PHASE III: OPTIMIZE & SECURE** the new environment. Optimize performance for maximum throughput per node. Manage and secure all of your data with enterprise-class tools.

ONE FRAMEWORK, BLAZING PERFORMANCE, IRON-CLAD SECURITY, DISRUPTIVE ECONOMICS

- Identify data & workloads most suitable for offload
- Focus on those that will deliver maximum savings & performance
- Access & move virtually any data to Hadoop with one tool
- Easily replicate existing workloads in Hadoop using a graphical user interface
- Deploy on premises & in the cloud
- Optimize the new environment
- Manage & secure all your data with business-class tools

ABOUT US

Syncsort provides fast, secure, enterprise-grade software spanning Big Data solutions from Hadoop to Big Iron on mainframes. We help customers around the world collect, process and distribute more data in less time, with fewer resources and lower costs. 87 of the Fortune 100 companies are Syncsort customers, and Syncsort's products are used in more than 85 countries to offload expensive and inefficient legacy data workloads, speed data warehouse and mainframe processing, and optimize cloud data integration. To discover the power of our products, visit **www.syncsort.com/try**

syncsort

LIKE THIS? SHARE IT!