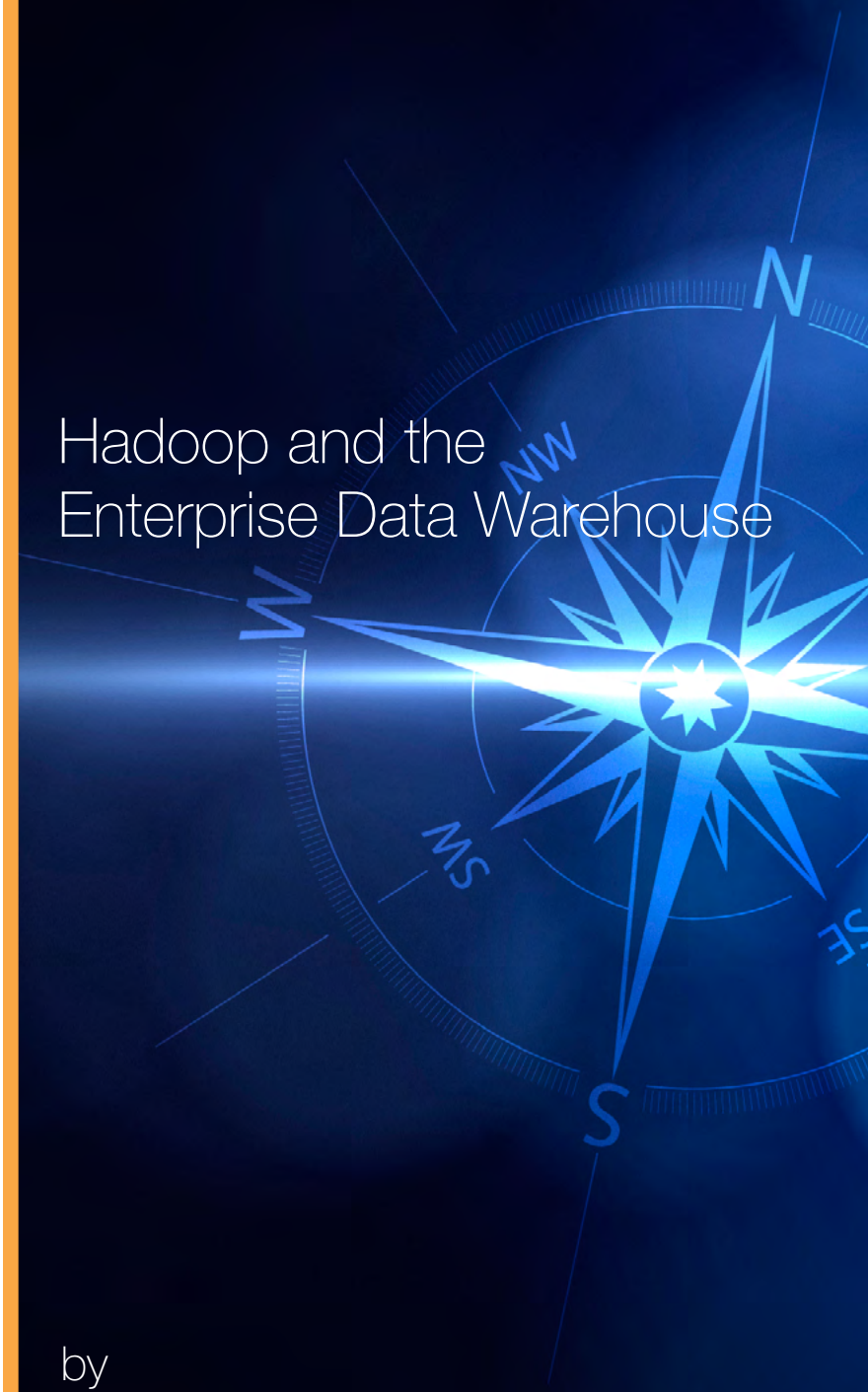
big **DATA**

A NON-GEEK'S BIG DATA PLAYBOOK

a SAS Best Practices white paper

Hadoop and the Enterprise Data Warehouse



by
Tamara **DULL**

sas best practices
THOUGHT PROVOKING BUSINESS

table of **CONTENTS**

This Big Data Playbook demonstrates in six common "plays" how Apache Hadoop supports and extends the EDW ecosystem.

INTRODUCTION

You: "No, we don't need Hadoop. We don't have any big data."

Big Data Geek: "Wah wah wawah."

You: "Seriously, even if we did have big data, that would be the least of our worries. We're having enough trouble getting our 'small' data stored, processed and analyzed in a timely manner. First things first."

Big Data Geek: "Wah wah wah wawah wah wah wawah."

You: "What's wrong with you? Do you even speak English?"

Big Data Geek: "Wawah."

You: "Right. Anyway, the other challenge we've been having is that our data is growing faster than our data warehouse and our budget. Can Hadoop help with that?"

Big Data Geek: "Wah wah wawah wah wah wah wah wawah wah wawah!"

You: "Whatever, dude."

We've all been there. We're having lunch with a few colleagues, and the conversation shifts to big data. And then Charlie Brown's teacher shows up. We try not to look at our watches.

If you're interested in getting past the "wah wah wawah" of big data, then this white paper is for you. It's designed to be a visual playbook for the non-geek yet technically savvy business professional who is still trying to understand how big data, specifically Hadoop, impacts the enterprise data game we've all been playing for years.

Specifically, this Big Data Playbook demonstrates in six common "plays" how Apache Hadoop, the open source poster child for big data technologies, supports and extends the enterprise data warehouse (EDW) ecosystem. It begins with a simple, popular play and progresses to more complex, integrated plays.

TERMS & DIAGRAMS

used in this playbook

This reference section defines the key terms used in the play diagrams.

The Enterprise Data Warehouse

Figure 1 portrays a simplified, traditional EDW setup:

Each object in the diagram represents a key component in the EDW ecosystem:

- **Structured data sources:** This is the data creation component. Typically, these are applications that capture transactional data that gets stored in a relational database. Example sources include: ERP, CRM, financial data, POS data, trouble tickets, e-commerce and legacy apps.
- **Enterprise data warehouse (EDW):** This is the data storage component. The EDW is a repository of integrated data from multiple structured data sources used for reporting and data analysis. Data integration tools, such as ETL, are typically used to extract, transform and load structured data into a relational or column-oriented DBMS. Example storage components include: operational warehouse, analytical warehouse (or sandbox), data mart, operational data store (ODS) and data warehouse appliance.
- **BI/analytics:** This is the data action component. These are the applications, tools and utilities designed for users to access, interact, analyze and make decisions using data in relational databases and warehouses. It's important to note that many traditional vendors have also extended their BI/analytics products to support Hadoop. Example applications include: operational reporting, ad hoc queries, OLAP, descriptive analytics, predictive analytics, prescriptive analytics and data visualization.

Big Data and Hadoop

Figure 2 represents a simple standalone Hadoop setup:

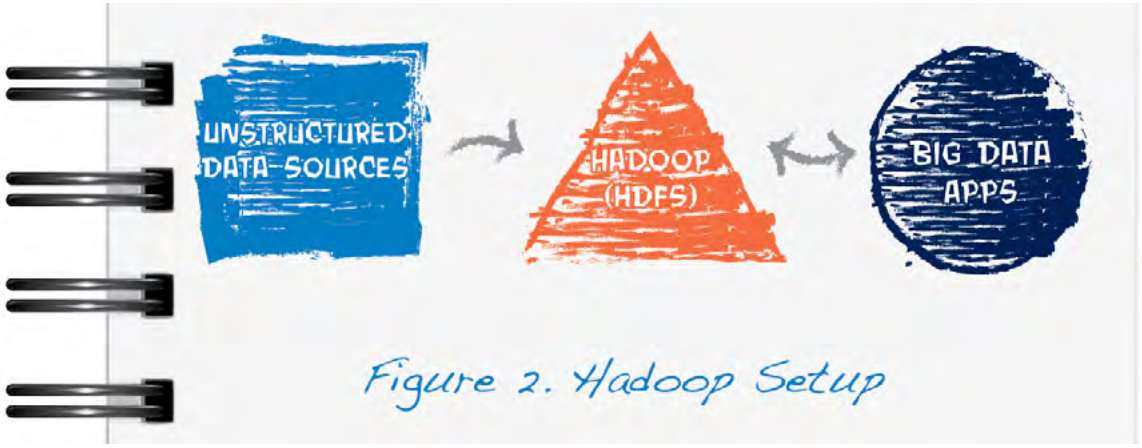


Figure 2. Hadoop Setup

Each object in this diagram represents key Hadoop-related components:

- **Unstructured data sources:** This is the data creation component. Typically, this is data that's not or cannot be stored in a structured, relational database. Includes both semi-structured and unstructured data sources. Example sources include: email, social data, XML data, videos, audio files, photos, GPS, satellite images, sensor data, spreadsheets, web log data, mobile data, RFID tags and PDF docs.
- **Hadoop (HDFS):** The Hadoop Distributed File System (HDFS) is the data storage component of the open source Apache Hadoop project. It can store any type of data – structured, semi-structured and unstructured. It is designed to run on low-cost commodity hardware and is able to scale out quickly and cheaply across thousands of machines.
- **Big data apps:** This is the data action component. These are the applications, tools and utilities that have been natively built for users to access, interact, analyze and make decisions using data in Hadoop and other nonrelational storage systems. It does not include traditional BI/analytics applications or tools that have been extended to support Hadoop.

Not represented directly in Figure 2 is MapReduce, the resource management and processing component of Hadoop. MapReduce allows Hadoop developers to write optimized programs that can process large volumes of data, structured and unstructured, in parallel across clusters of machines in a reliable and fault-tolerant manner. For instance, a programmer can use MapReduce to find friends or calculate the average number of contacts in a social network application, or process web access log stats to analyze web traffic volume and patterns.

Another benefit of MapReduce is that it processes the data where it resides (in HDFS) instead of moving it around, as is sometimes the case in a traditional EDW system. It also comes with a built-in recovery system – so if one machine goes down, MapReduce knows where to go to get another copy of the data.

Although MapReduce processing is lightning fast when compared to more traditional methods, its jobs must be run in batch mode. This has proven to be a limitation for organizations that need to process data more frequently and/or closer to real time. The good news is that with the release of Hadoop 2.0, the resource management functionality has been packaged separately (it's called YARN) so that MapReduce doesn't get bottlenecked and can stay focused on what it does best: processing data.

With the release of Hadoop 2.0, MapReduce doesn't get bottlenecked and can stay focused on what it does best: processing data.

PLAY 1

stage structured data

Use Hadoop as a data staging platform for your EDW.

With growing volumes of data and increasing requirements to process and analyze data even faster, organizations are faced with three options these days:

1. To add more hardware and/or horsepower to their existing EDW and operational systems.
2. Consider alternative ways to manage their data.
3. Do nothing.

While option 1 is viable yet expensive, and option 3 could be the kiss of death for some, option 2 is where Hadoop steps in.

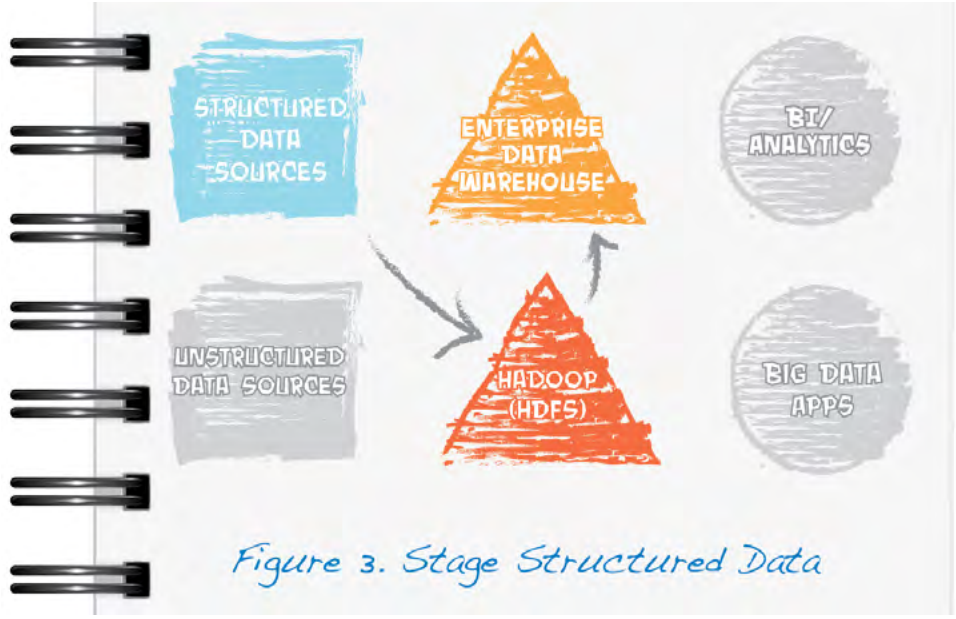


Figure 3. Stage Structured Data

With growing volumes of data and increasing requirements to process and analyze data even faster, organizations are faced with three options these days.

Consider this: What if you used Hadoop as a data staging platform to load your EDW? You could write MapReduce jobs to bring the application data into the HDFS, transform it and then send the transformed data on its way to your EDW. Two key benefits of this approach would be:

- Storage costs. Because of the low cost of Hadoop storage, you could store both versions of the data in the HDFS: the *before* application data and the *after* transformed data. Your data would then all be in one place, making it easier to manage, reprocess (if needed) and analyze at a later date.
- Processing power. Processing data in Hadoop frees up EDW resources and gets data processed, transformed and into your EDW quicker so that the analysis work can begin.

Back in the early days of Hadoop, some went so far as to call Hadoop the "ETL killer," putting ETL vendors at risk and on the defensive. Fortunately, these vendors quickly responded with new HDFS connectors, making it easier for organizations to optimize their ETL investments in this new Hadoop world.

If you're experiencing rapid application data growth and/or you're having trouble getting all your ETL jobs to finish in a timely manner, consider handing off some of this work to Hadoop – using your ETL vendor's Hadoop/HDFS connector or MapReduce – and get ahead of your data, not behind it.

If you're experiencing rapid application data growth, consider handing off some of this work to Hadoop.

PLAY 2

process structured data

Use Hadoop to update data in your EDW and/or operational systems.

Contrary to popular belief, you don't need "big" data to take advantage of the power of Hadoop. Not only can you use Hadoop to ease the ETL burden of processing your "small" data for your EDW (see Play 1), you can also use it to offload some of the processing work you're currently asking your EDW to do.

Take, for example, social networks like Facebook and LinkedIn that maintain a list of mutual friends and contacts you have with each of your connections. This mutual list data is stored in a data warehouse and needs to be updated periodically to reflect your current state of connections. As you can imagine, this is a data-heavy, resource-intensive job for your EDW – a job that could easily be handled by Hadoop in a fraction of the time and cost.

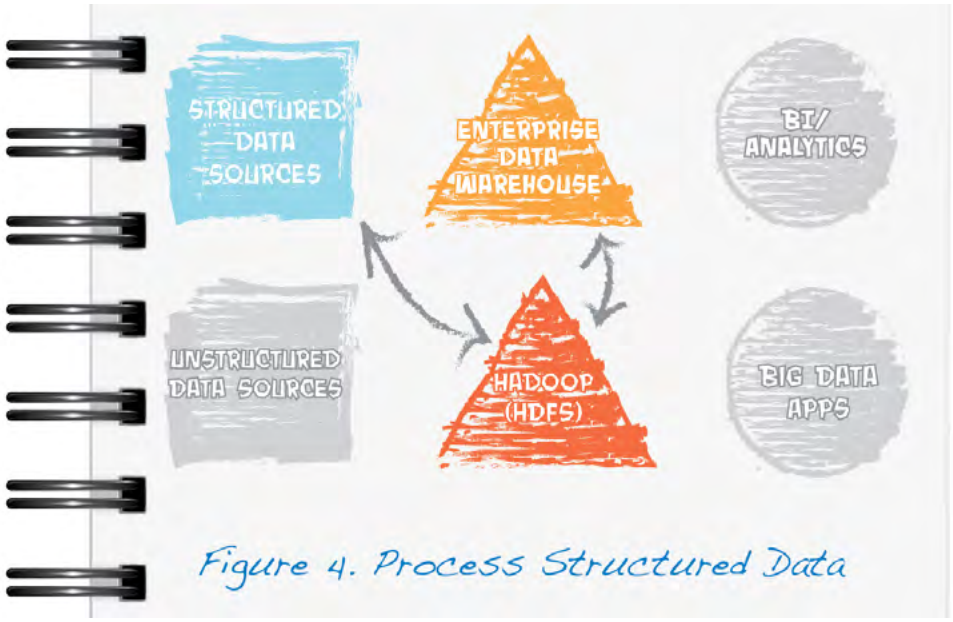


Figure 4. Process Structured Data

Figure 4 shows this play in action: Send the data to be updated to Hadoop, let MapReduce do its thing, and then send the updated data back to your EDW. This would not only apply to your EDW data, but also any data that is being maintained in your operational and analytical systems.

Take advantage of Hadoop's low-cost, high-speed processing power so that your EDW and operational systems are freed up to do what they do best.

Contrary to popular belief, you don't need "big" data to take advantage of the power of Hadoop.

PLAY 3

process non-integrated & unstructured data

Use Hadoop to take advantage of data that's currently unavailable in your EDW.

This play focuses on two categories of data: (1) structured data sources that have not been integrated into your EDW and (2) unstructured data sources. More generally, it's any data that's currently not part of your current EDW ecosystem that could be providing additional insight into your customers, products and services.

Notwithstanding, this raises a higher-level, more philosophical question: *Is all data valuable and should our goal be to capture all of it so that we can process, analyze and discover greater insights in it?* At many companies, such discussions are informed by a structured process for data governance, which is increasingly incorporating policy-making processes specific to Hadoop-resident data. It's important to note that even if the answer is "yes, collect it all," Hadoop is well-suited to help get us to this desired state.

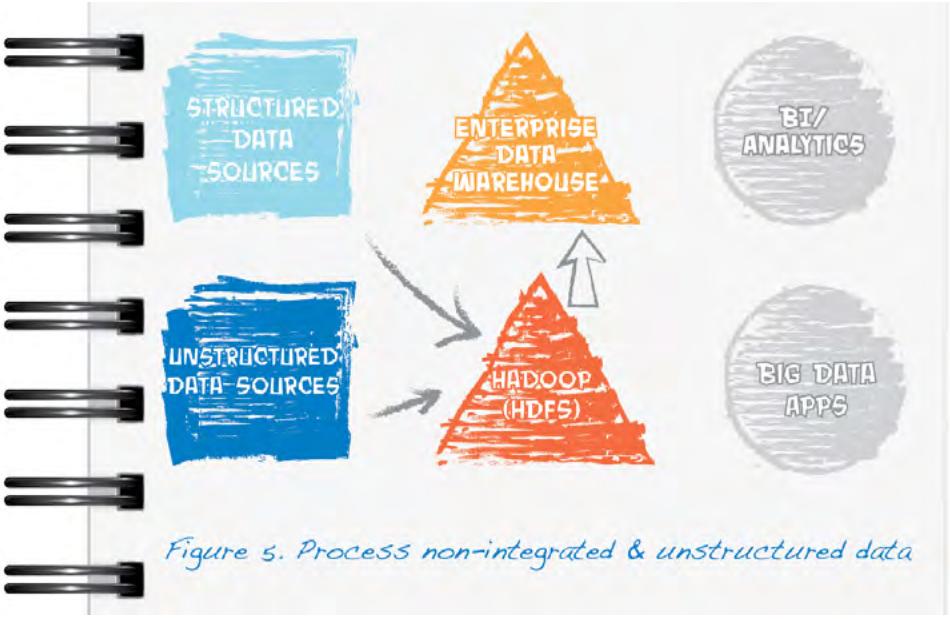


Figure 5. Process non-integrated & unstructured data

> Is all data valuable and should our goal be to capture all of it so that we can process, analyze and discover greater insights in it?

In Figure 5, data is coming into Hadoop (HDFS) from both structured and unstructured data sources. With the data in the HDFS, you have a couple of options:

- Process and keep the data in Hadoop (HDFS). You can then analyze the data using big data apps or BI/analytics tools. (See Play 6 for more information.)
- Process and store the data in Hadoop (HDFS) and, optionally, push relevant data into your EDW so that it can be analyzed with existing data. Be aware that not all unstructured data can or should be structured for the EDW.

Because Hadoop can store any data, it complements the EDW well. For the data your EDW cannot or doesn't handle well, you can use Hadoop to pick up the slack.

For the data your EDW cannot or doesn't handle well, you can use Hadoop to pick up the slack.

PLAY 4

archive all data

Use Hadoop to archive all your data on-premises or in the cloud.

This play is straightforward and one of the more popular plays. Since Hadoop runs on commodity hardware that scales out easily and quickly, organizations are now able to store and archive a lot more data at a much lower cost. To reduce costs even more, organizations can also archive their data in the cloud, thus freeing up more resources in the data center.

This is good news for IT, but it should also be music to the business professional's ears. No longer does data need to be destroyed after its regulatory life to save on storage costs. No longer does the business analyst or data scientist need to limit his data analysis to the last three, five or seven years. Because Hadoop is open source software running on commodity hardware and because performance can outstrip that of traditional databases, decades of data can now be stored more easily and cost-effectively.

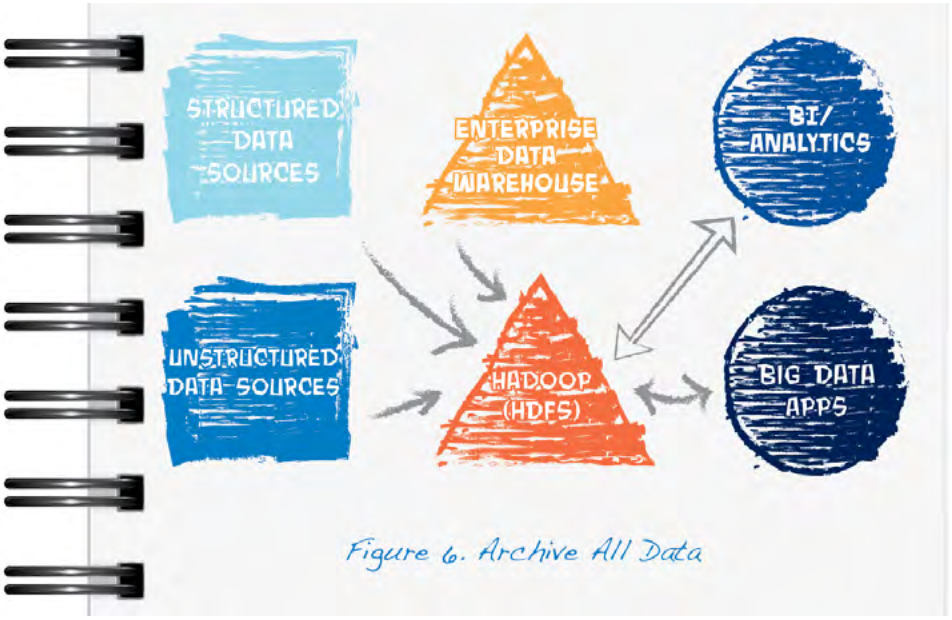


Figure 6. Archive All Data

As demonstrated in Figure 6, archived data in Hadoop can be accessed and analyzed using big data tools or, alternatively, using traditional BI/analytics tools that have been extended to work with Hadoop. Ideally, the best tools to use will be those that are most familiar to the data professionals and have been designed to handle the volume and variety of archived data.

No longer does the business analyst or data scientist need to limit his data analysis to the last three, five or seven years.

PLAY 5

access all data via the EDW

Use Hadoop to extend your EDW as the center of your organization's data universe.

This play is geared towards organizations that want to keep the EDW as the de facto system of record – at least for now. Hadoop is used to process and integrate structured and unstructured data to load into the EDW. The organization can continue using its BI/analytics tools to access data in the EDW, and alternatively, in Hadoop.

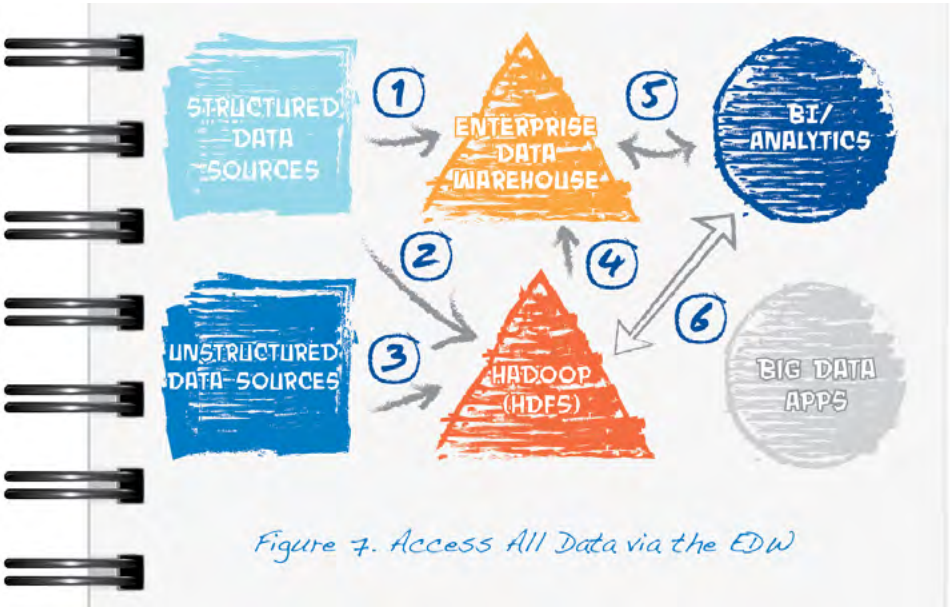


Figure 7. Access All Data via the EDW

Play 5 is geared towards organizations that want to keep the EDW as the de facto system of record.

Figure 7 demonstrates how this works:

1. This step focuses on structured data that meets two criteria: (1) Data that is currently being integrated (or can be integrated) into the EDW via ETL tools and (2) data that does not need to be integrated with other non-integrated or unstructured data sources (see Play 3).
2. This is structured data you want to integrate with other data, structured or unstructured. The best place to bring it together is Hadoop.
3. This is any unstructured data that you want to process and possibly integrate with other structured or unstructured data sources. This data can be stored in Hadoop in its raw state.

4. As outlined in Play 3, this is data that Hadoop has processed and integrated, and can be loaded into the EDW. Given that this play is about making the EDW the central player, figuring out which data goes to the EDW will be key.
5. All EDW data – from structured data sources and Hadoop – can now be accessed and analyzed via BI/analytics tools.
6. Alternatively, BI/analytics tools that have been extended to work with Hadoop can access the data in Hadoop directly. Be aware that the Hadoop view, by design, will not be a complete view of the data, given that the EDW is the de facto system of record in this play.

Be aware that the Hadoop view, by design, will not be a complete view of the data, given that the EDW is the de facto system of record in this play.

Organizations have been spending years building up and out their EDW ecosystem, and along comes Hadoop to upset the apple cart.

PLAY 6

access all data via Hadoop

Use Hadoop as the landing platform for all data and exploit the strengths of both the EDW and Hadoop.

This play is a paradigm shift in data management. Organizations have been spending years building up and out their EDW ecosystem, and along comes Hadoop to upset the apple cart. This play is focused on exploiting the strengths of both technologies, EDW and Hadoop, so that organizations can extract even more value and insight from one of their greatest strategic assets – data.

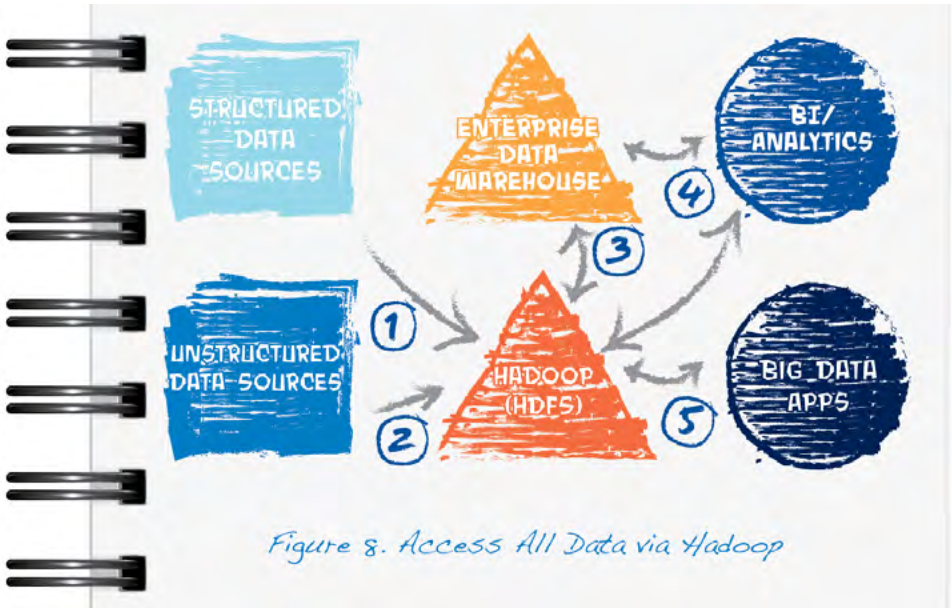


Figure 8. Access All Data via Hadoop

Figure 8 shows how it could work:

1. Structured data is initially processed and stored in Hadoop. It is no longer loaded directly into the EDW via ETL tools.
2. Unstructured data is collected, processed and integrated in Hadoop with other structured and unstructured data sources as needed.
3. Hadoop then becomes the primary source to load clean, integrated data into the EDW. The EDW also uses Hadoop to keep its data updated (see Play 2).
4. BI/analytics tools are then used to access and analyze data in both the EDW and Hadoop, depending on the business requirement.
5. Big data apps can and should be used to access and analyze data in Hadoop, now the repository for all data.

One advantage of capturing data in Hadoop is that it can be stored in its raw, native state. It does not need to be formatted upfront as with traditional, structured data stores; it can be formatted at the time of the data request. This process of formatting the data at the time of the query is called "late binding" and is a growing practice for companies that want to avoid protracted data transformation when loading the data. Late binding ensures that there is context to the data formats depending on the data request itself. Thus, Hadoop programmers can save months of programming by loading data in its native state.

Even though Figure 8 makes this data paradigm shift look simple, the road to get there will not be easy. Big data technologies, including Hadoop, are relatively new and are still maturing, especially for the enterprise market. Additionally, you can expect to see significant growth and maturation in the development of big data applications. Given that data creation is growing exponentially, it follows that we'll need better applications and tools to consume data and make it more meaningful, insightful and useful.

Big data technologies, including Hadoop, are relatively new and are still maturing, especially for the enterprise market.

CONCLUSION

Don't fall into the trap of believing that Hadoop is a big-data-only solution.

You: "I always thought you needed big data if you were going to use Hadoop. That's clearly not the case."

Big Data Geek: "Wah wah wawah."

You: "That's exactly right, and that's what I told my boss."

Big Data Geek: "Wah wah wah wawah wah wah wawah."

You: "Really? I'm going to check with IT and see how many years of data we're currently archiving. Maybe we can bump it up if we implement Hadoop."

Big Data Geek: "Wawah."

You: "Good point. I think we're currently archiving our data on-premises. Maybe we should look into moving some of this data off-site into the cloud."

Big Data Geek: "Wah wah wawah wah wah wah wah wawah wah wawah!

You: "Whatever, dude."

Don't fall into the trap of believing that Hadoop is a big-data-only solution. What's important for, indeed incumbent on, the tech-savvy business professional is to identify the improvement opportunities that Hadoop can help address. Often, technical practitioners and programmers are so focused on getting the data in that they lose sight of the promised efficiencies, not to mention business-specific constraints that might be lifted as a result of using Hadoop.

Whether or not you speak open source and whether or not you're database-fluent, you now have the business perspective to identify the optimal play(s) from the Big Data Playbook. In so doing, you can and will be the bridge between the problem and the solution.

[you have been reading a SAS Best Practices white paper. thank you.]

about the author

TAMARA DULL has over 25 years of technology services experience, with a strong foundation in data analysis, design and development. This has served as a natural bridge to Tamara's current role as SAS' thought leadership guru of big data, guiding the conversation on this emerging trend as she examines everything from basic principles to architectures and delivery best practices.

A pioneer in the development of social media and online strategy, Tamara has established dynamic marketing efforts and fostered robust online collaboration and community interaction. At Lyzasoft, she helmed a major release of the startup company's enterprise collaboration software suite. Tamara also established an online community in her role as the co-founder of Semper Vita, a non-profit charity website. Her role as the VP of Marketing at Baseline Consulting saw her provide strategic leadership and online media expertise for key marketing efforts and branding.

sas best practices
THOUGHT PROVOKING BUSINESS

SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513-2414
USA
Phone: 919-677-8000
Fax: 919-677-4444