White Paper

The Economic Advantages of Business Intelligence on Big Data

By Nik Rouda, ESG Senior Analyst; and Mike Leone, Senior ESG Lab Analyst

April 2017

This ESG White Paper was commissioned by AtScale and is distributed under license from ESG.

Contents

Introduction: Extracting Insights from Data, but at What Cost?

As companies look for more cost-effective ways to meet enterprise requirements for data discovery, organization, analytics, and reporting in the era of big data, the roles of traditional data warehouses (DW) and business intelligence (BI) are being reconsidered, driven largely by changing economics. It's not that the traditional approaches to extracting insights from data don't work—in general, they work very well. With decades to mature, these solutions are for the most part reliable, well defined, richly featured, and able to support large numbers of concurrent users. Yet pressures of our modern era are forcing many businesses to rethink their value.

The issue is that more data of greater variety challenges the scalability of traditional environments, especially in terms of economics. The cost of storage is directly linked to the volume of data, and while the quality of traditional data warehouses is generally high, they are also very expensive. Similarly, the need to analyze all of the data in a reasonable amount of time requires increasing amounts of server resources (particularly processors and memory), which, again, increases costs disproportionately. Scaling up to bigger and badder hardware on-premises is only sustainable to a point.

More subtly, if no less importantly, are the soft costs of managing traditional environments, where the rigid design of traditional data platforms doesn't jive well with a wide variety of unstructured data types. Extract, transform, load (ETL) of data soon becomes resource-intensive and unduly expensive. Performing exploration and discovery on data of unknown value takes a lot of time and tools. The human effort needed either restricts the number of questions that can be answered, or alternatively, has a real price tag in the man-hours required to keep up with business demands. All of these considerations support the need for a fresh approach.

> If nothing else, the era of big data is defined by the desire to handle larger volumes of data and to do so more cost-efficiently.

Big data, including Hadoop and data lakes, is rapidly gaining ground against traditional approaches for particular use cases, allowing more data to be kept and analyzed at significantly lower prices. Many are also going off-premises to leverage public cloud services to accelerate their big data on-ramp and reduce costs of infrastructure at scale. Now, even as big data provides welcome disruption, it isn't going to displace everything, not in the near term, anyway. Few experts believe traditional BI and data warehouses are dead; rather, they think that these approaches can be complemented nicely by big data technologies. For many, the goal is "better together," as shown in Figure 1, where 36% of respondents to a 2016 ESG research study reported that they expect to run their data warehouse and big data side by side, each doing what it does best, and together providing a better blend of costs.[1]

Yet Figure 1 shows that an attitude shift is taking place in the market, too. One significant shift is how enterprise data flows in a combined environment. Big data is not just about offloading low-value data warehouse functions (like ETL) to data lakes, where data is moved into Hadoop. In a striking reversal, many leading organizations we work with are looking to land all data in the data lake first, discovering relevant uses for it through exploratory BI in place. Some who tightly hold to traditional ways still migrate highly intensive or tightly defined subsets into the data warehouse. While this secondary strategy for data management is appealing as an attempt to leverage existing tools and skills, moving data more than once still introduces more complexity than anyone wants.

[1] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends: Redux*, July 2016.

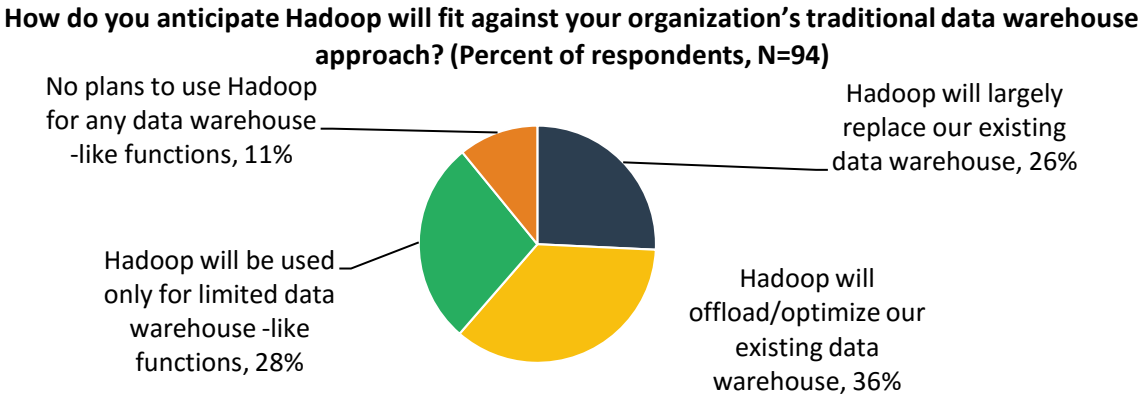
Clearly, a big-data-centric or fully big data approach is gaining popularity. For many, the determining factor in which path to take forward will be a calculation of value of data insights, which explicitly compares capabilities against costs.

62% expect big data to have a significant impact on their traditional data warehouses.

Altogether, 62% of the respondents to our 2016 survey expected big data to have a significant impact on their traditional data warehouses.

As this transition from traditional data warehouse to big data progresses, one critical element will be ensuring that analysts and knowledge workers are still able to do BI on the new platforms.

Figure 1. Anticipated Fit for Hadoop Against Traditional Warehouse Approach



Source: Enterprise Strategy Group, 2017

What's So Hard About BI on Big Data and Data Lakes?

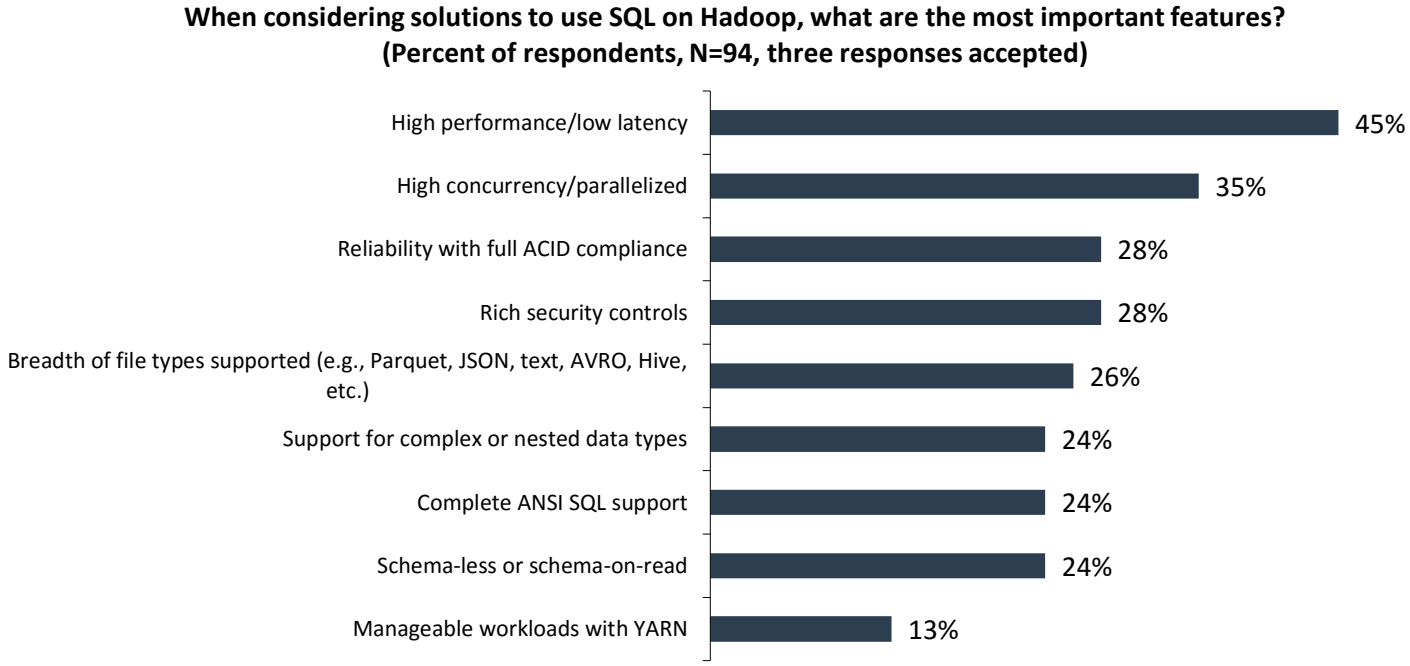
Whether you choose a pure big data or a hybrid environment of data warehouse and data lake, it will still require an ability to perform BI-style analytics effectively. Many innovators who jumped in early have discovered that doing BI on big data platforms (including Hadoop) has proven trickier said than done. There are many flavors of SQL on big data engines available today (such as Impala, Hive, Drill, and Presto), but organizations often struggle to match the right engine across the many functional requirements of the business.

A recent ESG research study on requirements for SQL on big data shows performance and concurrency as top demands, but respondents also include a number of other features needed to match the quality expectations of those used to doing analytics on traditional data warehouse environments (see Figure 2).[2]

While organizations could develop their own custom solutions to meet some of the requirements around SQL on Hadoop and subsequently BI on big data, these workarounds bring real costs. An enterprise-grade BI-on-big-data solution should address *all* of the concerns listed in Figure 2 and offer more confidence that this approach is mature enough to be accepted by the business as an alternative to traditional BI and data warehouses.

[2] ibid.

Figure 2. Most Important Features for Solutions to Use SQL on Hadoop



Source: Enterprise Strategy Group, 2017

Companies looking to drive a BI-on-big-data strategy tend to overlook a number of key challenges, and sometimes these challenges become unavoidable problems and erode value if left unaddressed. To be successful with BI on big data, you should consider:

- Moving big data is unwieldy. The reality of big data is that there's too much data—and that it's too varied and coming too fast—to easily do ETL. Data movement and processing can delay the start of analytics (sometimes by weeks), increase data redundancy, and waste time, effort, and money.
- Unfamiliar interfaces make BI on big data harder to use. Big data may revolutionize the world of data, but business analysts shouldn't have to learn all new BI tools. If you don't enable analysts and business users to access the BI platforms they already know, adoption rates of the new tools will be slower and lower than desired.
- Standards are needed. While the pace of change in big data is impressive, too much innovation can lead to chaos. Organizations without common standards for their applications and analytics are going to use data inefficiently, independent of any particular source. Without a standard semantic layer and standard vocabulary for analytics, you'll put significant effort and cost into managing data.
- Secure access to data is essential. As already noted, data needs definitions, users need roles, and sensitive data must be identified and controlled accordingly. Without these functions, data is effectively unusable and/or insecure. Basic big data platforms generally are still lacking sufficient security and governance policies and potentially introduce expensive breaches.

As you evaluate potential products and services to help you implement your BI-on-big-data initiative, it's critical that you compare how each offering addresses the concerns listed here.

How To Deliver Effective BI on Big Data

Organizations will want the best of both big data and BI worlds, and the good news is that the problems outlined above can be overcome with the right solution. To satisfy everyone, IT must have the control and quality needed for the data infrastructure, and the business must have timely access to data and familiar BI tools to make them successful.

An ideal solution that brings this balance would deliver:

- In-place BI on *all* data in the data lake. Once data lands in a data lake, immediate BI and analytics could be performed right there, without moving the data out again. The value and accuracy of insights would be increased by effectively and efficiently using all available relevant data, based on simple discovery and pervasive cataloging. More insights lead to more ROI for analytics for the business.

- The ability to use existing BI and analytics tools. The obvious advantage to reusing already mastered skills and already deployed tools is that no learning or change of established practices would be required. Existing BI platforms and analytics languages could be kept, without incurring any additional cost or effort, greatly simplifying matters for both IT and business analysts. Knowledge workers can just go find insights with the tools they already know, and be productive on the broader datasets immediately, increasing the "time value" of their work.

- Upheld and reinforced standards. A universal dataset should be complemented by a flexible object representation with semantics to support all usage and analytics tools. There would be no ongoing effort required for repeatedly adapting data and metadata for different sources and analytics engines. Instead, data could be leveraged quickly and consistently across all data management platforms. This improves accuracy while at the same time reducing complexity and effort around data definitions.

- Quality and controls built in. Businesses won't have to reinvent and build their own frameworks, or invest in buying new utilities to provide necessary safety and security for their big data environments. Existing best practices will be preserved by default, giving more confidence and reducing risks.

Most importantly, each of these benefits will translate directly into economic advantages. Indeed, the only way the economics of big data to complement (or replace) data warehouses actually work is if these qualities can be assured!

Cost Benefits of a Modern Platform for Business Intelligence with Big Data

This section will help you identify and quantify both the hard and soft costs associated with data movement and massively parallel processing (MPP).

ESG analyzed the costs and value of leveraging a modern platform to perform business intelligence (BI) on big data as opposed to using several traditional approaches. These traditional approaches often include costly enterprise data warehouses (EDW), creating and rebuilding OLAP cubes, and using multiple ETLs to move data into a specific format or destination. And this is all done before custom BI tools can help with analysis and visualization.

The goal of ESG's analysis was to highlight how a flexible, scalable modern BI platform can help control costs, improve productivity, and reduce dependencies on proprietary, inefficient, expensive BI architectures and approaches. When evaluating the hard and soft costs of an infrastructure to handle BI on big data, ESG recommends focusing on four main areas:

- **Infrastructure costs:** Hardware for MPP, annual license fees for core software, and annual maintenance/support fees.
- **Labor costs:** Annual salaries of staff responsible for platform management and data administration.
- **Common BI tasks:** Time to load and design data, along with time to create/build OLAP cubes.
- **BI tool integration:** Time to modify semantic models based on a single BI tool.

Infrastructure Considerations

For Traditional Big Data: Traditional approaches to BI on big data generally consist of on-premises, non-Hadoop hardware and software that are usually proprietary. This fact alone tends to cause inflated costs for the data management side of things. On top of the hard costs, traditional infrastructure platforms tend to be inflexible, with limitations on scalability and interoperability. For example, if an initial deployment uses hardware from Vendor A, and more hardware is needed, the only options are buying more hardware from Vendor A, or putting additional investment in place to account for Vendor B being added to the infrastructure.

This leaves organizations with two options if they want to scale: higher cost from vendor lock-in or more complexity from a mixed-vendor infrastructure due to incompatibility between vendors. Further, using multiple platforms leads to multiple copies of the same datasets appearing in both vendors' platforms, leading to unnecessary storage consumption. Worse, one of those datasets may be more up to date than another. When combined, the inefficiencies, disconnects, and false or misleading insights can quickly drain budgets and hurt an organization's competitive advantage.

For Modern BI on Big Data: By turning to a modern approach for BI on big data, organizations can gain the immediate cost savings and flexibility associated with open source technology. As an example, if an organization leverages Hadoop as the core underlying big data hardware platform, it can use commodity servers to save on initial costs, while enabling freedom of choice when selecting the server vendor. This approach also enables organizations to eliminate silos of the same dataset that reside on multiple platforms. A single, scalable platform can be used to store all data, not only eliminating the need to move data, but also maximizing storage capacity efficiency with a single copy of data.

ESG analyzed the storage requirements of a medium-sized organization. We modeled a sample environment requiring a storage capacity range of 300–500 TB, and estimated the costs of both a traditional and modern approach. We based pricing on averages taken across multiple leading vendors in the BI space. We priced the traditional approach as if it were a traditional EDW, while we priced the modern approach using Hadoop. We used a node count of 40 because it is a good barometer for a generalized enterprise with a semi-large data environment. Table 1 (below) highlights basic costs assigned to hardware for MPP, annual software licenses, and annual maintenance support.

Table 1. Infrastructure Cost Comparison

Infrastructure Costs	Traditional Approach	Modern Approach	Savings (times less)
Hardware for MPP	$2,500,000	$800,000	3.1x
Annual License	$1,000,000	$250,000	4x
Annual Maintenance/Support	$500,000	$400,000	1.3x

The cost savings that can be achieved using a modern approach to BI for big data—versus a traditional MPP or EDW—are impressive. Hardware costs of a modern big data platform like Hadoop can be more than three times less. Further, annual software license fees are reduced by one-quarter of those of a traditional approach, while support and maintenance fees also show a savings of 20%.

Why This Matters

By leveraging a platform like Hadoop as the underlying hardware infrastructure for a modern approach to BI on big data, organizations no longer need expensive, specialized hardware to handle traditional BI storage requirements. Industry-standard servers can be used to easily build a large Hadoop cluster that can scale to meet dynamic business environment needs. Further, the money saved by using this modern approach can be repurposed for other strategic initiatives within the big data organization. And by using a single platform, organizations gain peace of mind knowing that no data movement is required to make data BI-ready and that the dataset they are looking at is the latest and greatest. Put it all together, and organizations can save hundreds of thousands of dollars in hardware, software, and support costs using a modern approach to BI on big data.

Labor Considerations

For Traditional Big Data: Labor costs can be broken into two categories: staff responsible for managing the platform, and data administrators responsible for common data workflow tasks. With a traditional big data management approach, organizations need employees with expertise on a specific vendor's platform to effectively manage the deployment.

Data administrators (or architects) are responsible for designing, creating, deploying, and managing the overall data architecture. Though some specialized skills are required, they do not need to fully understand the underlying hardware platform. Managing the data architecture is already a difficult task for data administrators, especially as more end-users want access to data, but imagine what happens when those additional end-users ask for access across multiple proprietary platforms using different BI tools. The potential for delays due to the constant movement of massive datasets and the data preparation required for a specific platform or BI tool have never been more apparent. In fact, hearing from one organization who currently uses a traditional BI approach, data had to be moved three or more times *before* the BI tool was even brought into the picture.

For Modern BI on Big Data: Using a modern approach for BI on big data, labor costs are easier to manage, even with a highly scalable environment consisting of hundreds of nodes. Whether scaling vertically or horizontally, the underlying infrastructure can easily support larger datasets without requiring more staff to manage it. Data administrators can rest

easy knowing that by using a single platform, they only need to develop a single set of tools to introduce higher levels of automation and operational efficiency. No data movement is required, and end-users can easily use their BI tool of choice to analyze a single copy of the same dataset. And the learning curve to become proficient at a modern approach can take as little as a single day if the existing big data staff already has experience in big data platforms and workflow management.

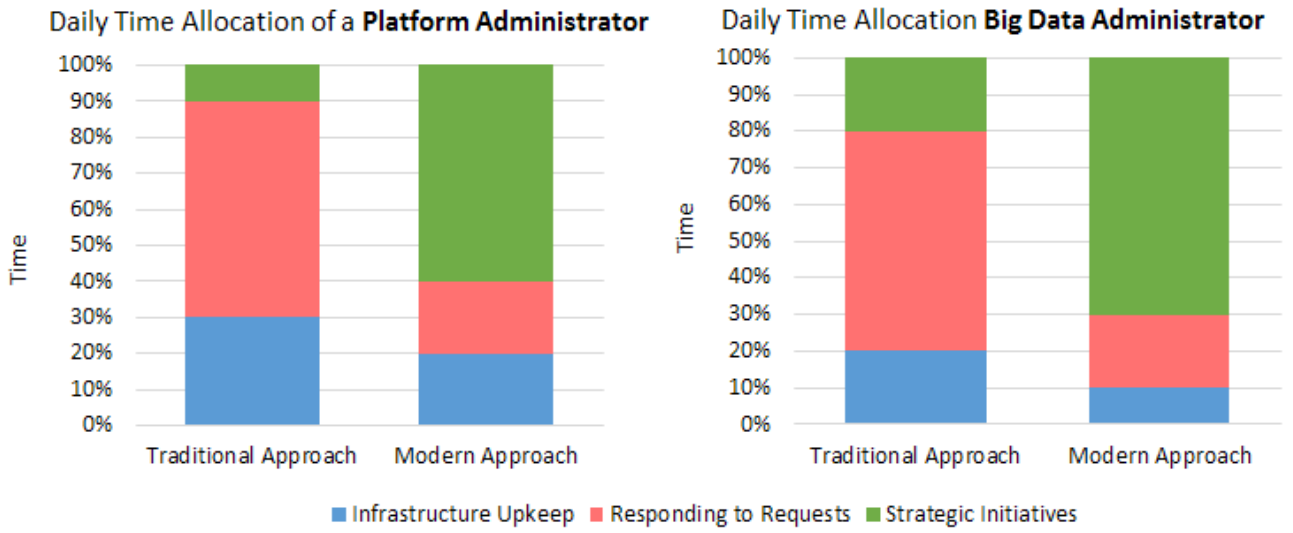
Figure 3 below shows how platform administrators' and data administrators' might spend their time during a typical day using a traditional approach for BI on big data, and compares it to how they would spend their time if they were using a modern approach. This theoretical example assumes that the hardware and software has already been deployed and estimates how administrators allocated their time throughout the day. We grouped the two administrators' main tasks into three buckets:

1. **Infrastructure upkeep, responding to requests, and strategic initiatives:** Infrastructure upkeep involves adding new components, updating existing components, and validating interoperability between components.

2. **Responding to requests:** This includes addressing end-user complaints related to performance or interoperability issues, while also moving data to its proper location based on end-user requests.

3. **Strategic initiatives:** This refers to the time the administrators spend adding more value to their organization.

Figure 3. Daily Time Allocation of Administrators



Source: Enterprise Strategy Group, 2017

Why This Matters

Though labor costs in traditional and modern approaches may appear the same on paper with the same employee count, the soft benefits that a modern platform offers organizations is unparalleled. ESG analyzed three key benefits for organizations taking a modern approach. Fewer infrastructure components mean less time on upkeep. In addition, this not only presents the opportunity to ensure better interoperability between the minimal components, but also, with no data movement, requires less time to deal with end-user requests or experienced issues. And finally, with less time being spent on infrastructure upkeep and responding to requests, a large chunk of time can be reallocated to more strategic initiatives within the organization. These initiatives can range from finding ways to improve total cost of ownership and return on investment using future improvements and investments, to solving more complex problems related to business efficiency to help drive more revenue.

Common BI Task Considerations

For Traditional Big Data: The rigidness of using a traditional approach for BI on big data is often rooted in its requirements for data to be structured for loading, designing, and modeling. This necessitates a high level of preparation and organization to ensure that data properly fits into a predefined schema (hence the term schema-on-write). And the time to structure data before loading grows nearly exponentially as data comes in from a growing number of disparate sources with different originating structures, whether from social media or weather feeds. ESG estimates the amount of time spent preparing data for loading to be as much as 30 hours per month for a single job. Imagine the time required for organizations that must complete the loading phase multiple times per month. And once that data is loaded, it usually has a limited amount of time that it can reside on the platform due to common capacity constraints with a traditional platform. In fact, ESG found that one actual organization that used a traditional approach could only hold a quarter's worth of data in its BI tool of choice, whether that was an OLAP cube or a smaller structure within the existing data warehouse that handles the exported data.

After loading data, organizations typically design or model it into a single entity, commonly known as an OLAP cube. Building an OLAP cube in a traditional environment can be a complex, time-consuming task using rigid constructs on proprietary hardware with specialized, unintuitive software. ESG estimated data modeling to take approximately 30 hours/month, while building a cube or preparing the cube object was set at 20 hours/month. Again, this is a conservative estimate, as ESG found one organization taking up to four days to refresh a single cube.

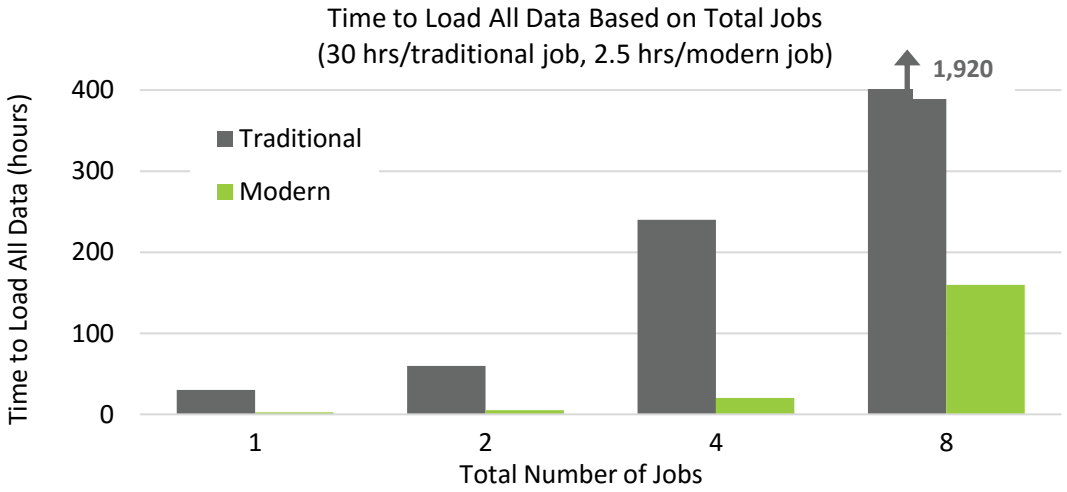
For Modern BI on Big Data: With a modern approach to BI on big data, using technology like Hadoop or other big data platforms, organizations can gain the ability to easily handle and store any type of data, whether structured or unstructured, without requiring major pre-work or adjustments for it to be stored. Further, that data can simply be read regardless of its form. Of course, metadata management is required to maintain consistency and data preparation is required to allow the BI tool of choice to read it, but ESG estimated that the impact to the overall time required to load and prep data on a modern platform was just 2.5 hours per month per job. To emphasize this time savings over a traditional approach, Figure 4 shows a calculation of how much time might be spent moving and loading data as the number of load jobs increases from one to eight over the course of one month. The modern approach saves an impressive amount of time—12 times that of a traditional approach—and this time savings is just for the loading data phase.

Note: For data modeling, ESG has seen that merging datasets typically uses similar processes and time as a traditional approach, and therefore we estimated the same amount of time for this task, 20 hours per month. Again, this is a

conservative estimate. In fact, many organizations that leverage a modern approach take far less time to model data. One organization saw data modeling times go from weeks to within a day when switching to a modern approach.

And finally, the creation of OLAP cubes using a modern platform provides organizations with familiar workflows to quickly and dynamically create virtual cubes using intelligent optimizations. This efficient manifestation of cube objects helps reduce traditional OLAP cube creation times from several weeks down to several hours. In fact, ESG estimated that creating OLAP cubes using a traditional approach requires approximately 20 hours per month, while a modern approach can complete this same process in one-fourth the time, saving 15 hours per month on cube creation and refresh alone.

Figure 4. Comparing Times to Load All Data



Source: Enterprise Strategy Group, 2017

Why This Matters

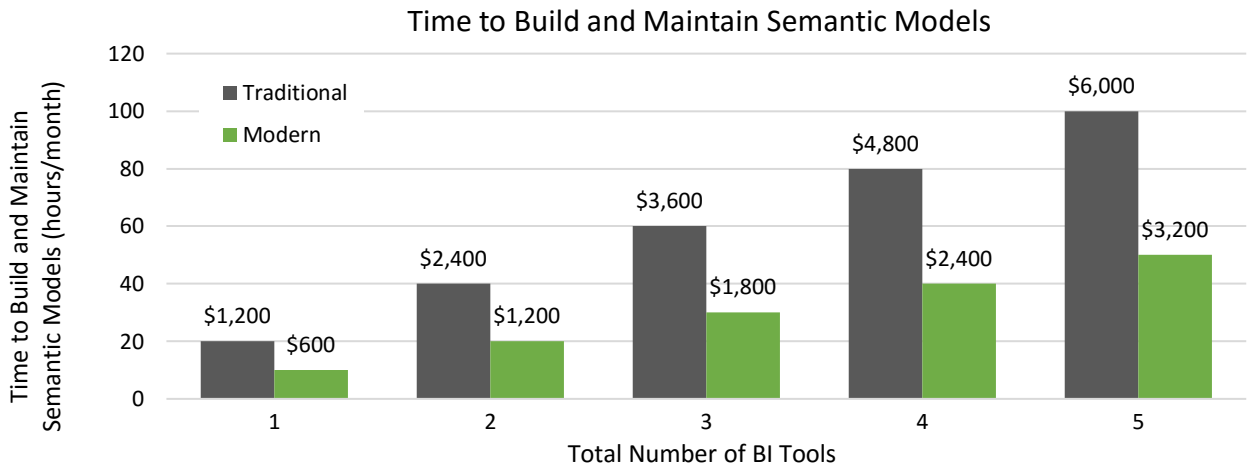
A modern approach to BI on big data enables organizations to load data significantly faster, without the need to modify anything related to the data. Whether the data is structured, semi-structured, or unstructured, modern big data platforms both on-premises and in the cloud can easily store it. This enables a reduction in the number of ETL jobs and data extracts, while allowing organizations to keep all their data together for potentially deeper insights. ESG analyzed the scalable time savings organizations could achieve as the number of data load jobs increased in a modern platform, and discovered that a modern approach would yield a 12 times savings when compared with a traditional approach. Further, organizations no longer must wait hours to rebuild BI-ready (OLAP) cubes. Using efficient and intelligent optimizations, a modern platform can complete the BI cube creation process in one-fourth the time of that of a traditional process, enabling faster data access and deeper insights for business analysts and organizations.

BI Tool Integration

For Traditional Big Data: Many organizations leverage multiple big data platforms from different vendors to handle traditional BI on big data. Each one of these vendors could very well have its own proprietary BI tool. On top of that, business analysts often use other third-party BI tools that are not necessarily affiliated with a specific vendor or technology. This creates a high level of complexity across the organization that requires specialists in each BI technology. Further, if data from one technology needs to be merged with data from another, the semantic model must be modified based on the BI tool being used to analyze the data. And in some cases, one specialized BI tool may not be able to handle data from a different vendor's technology. The standard amount of time enterprises spend building and maintaining semantic models was estimated at 20 hours per month per BI tool.

For Modern BI on Big Data: From a BI tool integration standpoint, modern BI for big data interfaces enable organizations to create or use a single, universal semantic model that can be leveraged across all BI tools, significantly reducing the semantic model change effort and making accurate data available to business analysts for analysis much faster. ESG calculated improved speed of data availability to save 50% of the time normally spent modifying semantic layers for each BI tool. And most modern interfaces use the latest and greatest technology like HTML5, giving organizations an improved look and feel that is simply easier to use and more intuitive than traditional interfaces.

Figure 5. Comparing Times to Load All Data



Source: Enterprise Strategy Group, 2017

Figure 5 highlights the time required and cost associated with building and maintaining BI tool semantic models using a traditional and a modern approach. By calculating $/hour of an employee responsible for completing this task, ESG could provide an even more compelling savings offered by a modern platform. Using the average salary of a big data engineer ($125,000) who works 40 hours a week, an estimate of $60/hour was calculated, allowing ESG to assign a value to each charted data point knowing that the process of building and maintaining semantic models was a manual process. As an example, using a traditional platform with three BI tools, we estimate that it would take 60 hours per month to build and maintain semantic models. The cost associated with that is $3,600/month or $43,200/year, which would consume nearly a third of one employee's time and salary just building and maintaining semantic models using a traditional platform.

Why This Matters

With a modern BI platform for big data, organizations can leverage existing investments in BI tools, providing users the freedom of choice to leverage existing BI tool skill sets. One semantic model can be used across all BI tools to access a universal, consistent, up-to-date data model. This not only enables faster access to accurate datasets for analysis, but also reduces the time spent by data administrators responding to constant requests to access the data. The value of efficient and accurate BI tool integration is paramount to gaining valuable insights as quickly as possible. Based on ESG's calculations, organizations can save 50% of cost and time creating and maintaining semantic models per BI tool when leveraging a modern platform for BI on big data.

The Bigger Truth

Big data is not just about the price advantage of open source software and commodity servers. If anything, the many hidden costs of managing a distributed infrastructure can be significant. Ultimately the activities and workflows of analytics on big data must be fast and efficient to deliver value, and cannot compromise on the many demands of the modern enterprise. If big data and data lakes can't meet these needs for insight, organizations will remain stuck with traditional data warehouses and won't reap the promised rewards. However, there are solutions that make BI on big data work for the enterprise. This paper has detailed both the logic and the numbers behind a modern approach to BI on big data.

ESG calculates that a modern approach to BI on big data results in overall infrastructure costs that are 64% lower than with a traditional approach. With labor, the value is found in how time is allocated, with a modern approach giving administrators an average of 65% of their time to focus on more strategic initiatives for the business, rather than spending the majority of their time performing ongoing platform maintenance and/or responding to constant end-user requests. Modern BI on big data allows organizations to complete common BI tasks, such as loading data, designing data, and building OLAP cubes 61% faster. Finally, BI tool integration done with a modern approach can yield significant time and cost savings, with organizations spending 50% less time and money building and maintaining semantic models than they would with a traditional approach. There is a better way to get BI insights from your data lake—it's high time to check it out!

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.

 www.esg-global.com

 contact@esg-global.com

 P. 508.482.0188