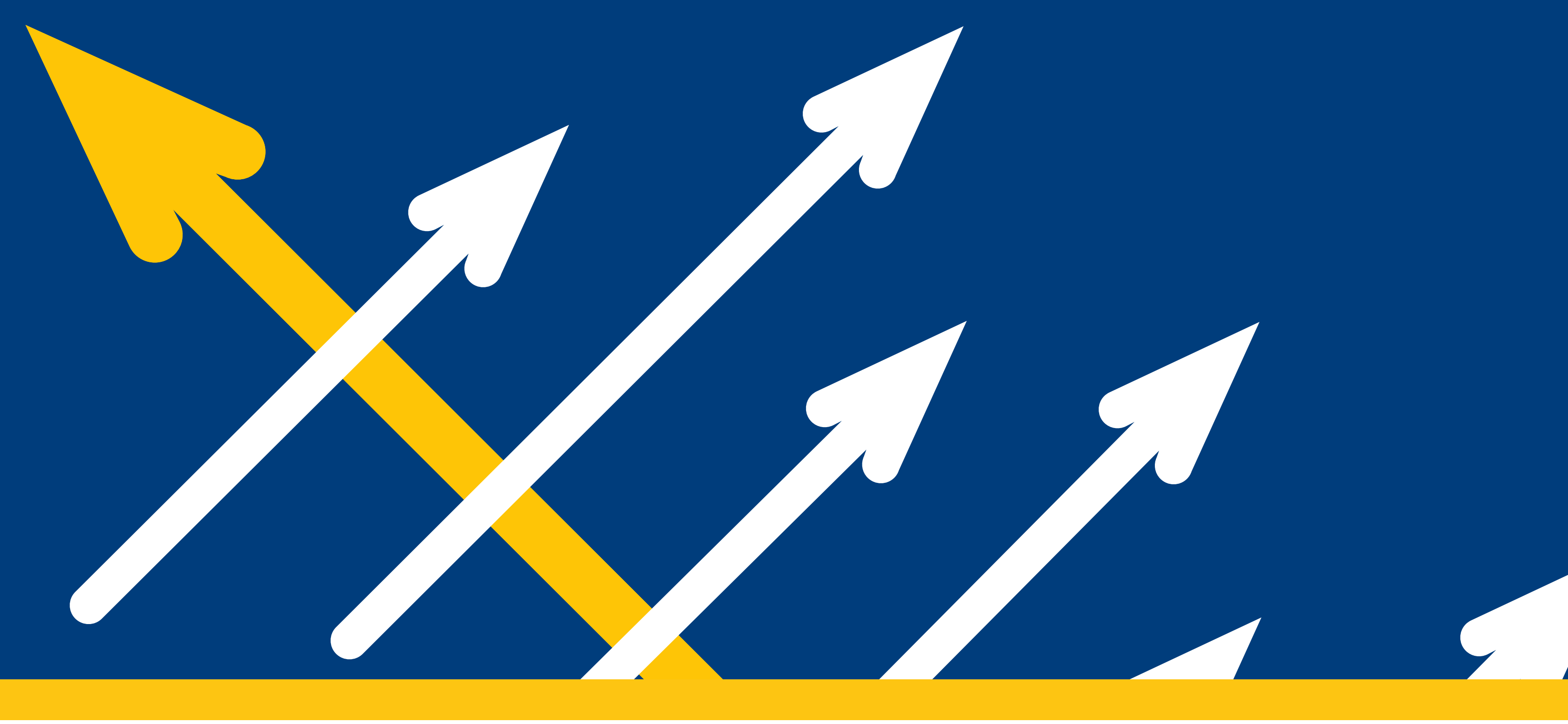
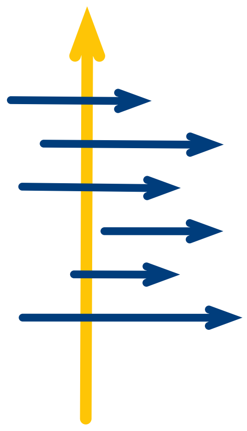


BI on Big Data: The Traps You Should Avoid

by Thomas W. Dinsmore



Data Is Not Enough

Congratulations! You invested time and money getting Hadoop up and running, Now, you can see a light at the end of the tunnel. Your data feeds are in place, and working properly.

Your team knows how to manage the cluster, scale it out, and make it work. Expert users in your organization work with Hive, Pig, Spark and other native Hadoop tools. It looks like they can produce real insight for your business.

However, your executive team isn't satisfied. They've read the hype around Hadoop. They read the latest analysts reports that pigeon hole your solution to the traditional "Native BI on Hadoop" tool.

Yet, you know that from your executive staff's perspective value means people using data – lots of people.

They want to see people using data every day:

- **Measuring results**
- **Forecasting sales**
- **Finding good prospects**
- **Optimizing campaigns**
- **Reducing churn**
- **Profiling customers**
- **Developing new customers**
- **Understanding costs**
- **...AND MUCH MORE**

A recent Forrester report predicted that 100% of enterprises will deploy Hadoop in the next 12 months.

The latest Hadoop maturity survey conducted by Cloudera, Hortonworks, MapR and AtScale highlights the connection that exists between business value and connecting business users to Big Data properly: organizations that provide business users with self-service access are nearly 50% more likely to realize tangible value.

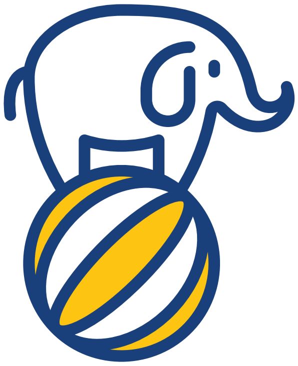
Yet, very few best practices exists. In the same survey, you'll find that 60% of respondents complain of a lack of self-service to Hadoop.

This paper is intended for technology executives who are eager to drive competitive advantage through Big Data, who want to move beyond the "Native BI on Hadoop" hype and who know that true value comes when every employee can use data, any data, in an unencumbered manner, with maximum agility and without compromising their corporate standards.

Traditional Business Intelligence (BI) software, such as SAP Business Objects or IBM Cognos, supports some of the above capabilities.

There is a problem, however: most of the leading BI tools do not work with Hadoop, or they perform poorly. In theory, BI tools can connect directly to structured data in Hadoop; in practice, however, this rarely works because BI users need data aggregated in specific ways to answer the questions they want to ask.

Ordinarily, a business user who wants to use tools like Tableau or Excel with Hadoop must first use Hive, Pig or Spark to create the desired aggregates – or get someone else to do so.



You have a short window.

In today's economy, executives have little patience with IT investments that don't deliver. They've seen slick demos and sales pitches from cloud-based vendors who promise insight in days. We both know that those claims are an exaggeration, but the message resonates because businesses need speed and agility today more than ever.

Unless you can get a lot of people using your Hadoop cluster, nobody will care how many nodes it has, or how many petabytes of data it can store.

Your leadership team will see the project as a white elephant.

A Quick Checklist: 5 things your company can't do without

To deliver on your corporate vision, your organization will need to deliver on key capabilities. Here is a quick checklist for what a mature organization is able to deliver on.

☑ ***Dimensional analysis.*** *Business analysts aggregate data in various ways for more meaningful information. The hierarchies they use for aggregation are unique to each business, and they can change. Users need the ability to define and modify hierarchies and use them to compute aggregate measures.*

You need this because your business measures performance at many different levels.

☑ ***Time-based measures.*** *Every business measures performance over time, and every company defines time periods in slightly different ways. Business users need the ability to define time-based hierarchies and fiscal periods; aggregate data by period; compare time periods, and adjust time-based measures for differences in the number of business days.*

You need this because your business measures performance over time.

☑ ***Simple drill-everywhere and drill-to-detail.*** *Business users ordinarily begin with aggregate measures, such as company-wide sales in a period, or total daily sales in a store. But if the high-level measure is surprising, they want to drill down to the details, on many different dimensions.*

You need this because you need to see the details when performance is surprising.

☑ ***Budgets and Benchmarks.*** *Actual performance data is never enough; business users want to compare performance with budget figures or other references. Budget information isn't always available in the source database, so the user must be able to supply the data separately, or quickly bring performance data into a tool like Microsoft Excel.*

You need this because you need to match performance and commitments.

☑ ***Sharable business definitions.*** *Assets such as hierarchies, organization charts, fiscal calendars and key performance indicators (KPIs) are unique to each business. Analysts need the ability to work with common standard definitions – and also to customize them when necessary.*

You need this because your business should measure performance consistently across units.

The Conventional Wisdom: Four Ways to Fail

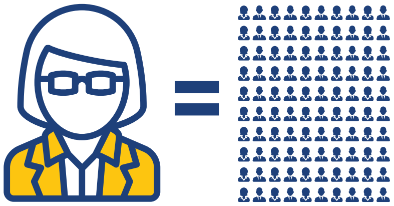
So, what do you do? According to the conventional wisdom (CW), there are four ways to deliver BI from your Hadoop data platform.

- **Copy the data to a relational database.**
- **Hire data scientists.**
- **Train business users to use tools like Hive, Pig, & Spark.**
- **Implement a native Hadoop BI platform.**

How good are these options? Let's take a look...

Copy the data.

Don't laugh. Traditional data warehouse vendors will tell you this is the smart thing to do: co-locate Hadoop with a data warehouse appliance, then use ETL to copy high-value summary data from one platform to another. Aside from the additional cost of maintaining multiple data platforms, this approach is bound to be a maintenance nightmare. Business hierarchies and metrics frequently change, so your team will be under constant pressure to modify the feeds. And, unless you are willing to invest in expensive real-time updates, this architecture adds another layer of latency to the data available to users. That won't sit well.



Hire data scientists

Data scientists have the skills needed to draw insights from Hadoop. "So, let's hire data scientists!" you say. It's not as easy as it sounds. Data scientists are like unicorns; The Chicago Tribune, McKinsey, Venture-Beat, and The Wall Street Journal all report on the shortage. The Harvard Business Review recommends that you stop looking, or lower your standards because qualified data scientists are hard to find. And even if you could build a team of data scientists, they won't stick around long if you set them to work running simple business queries. Data scientists can drive a lot of value for your company with advanced analytics, but you can't depend on them to meet your business intelligence needs.

Train business users

Every Hadoop distribution includes analysis tools like Apache Hive, Apache Pig, and Apache Spark. Depending on which distribution you use, you may have other tools for analytics, such as Apache Drill or Apache Phoenix. These are powerful tools; to use them, however, you must work in languages like Java, Scala or Python. Few business users know these, and most have neither the time nor the patience to learn. For most companies, this option simply is not viable.

Implement a "Native Hadoop BI Platform."

Several commercial vendors offer Business Intelligence (BI) software designed to run in Hadoop.

These software packages, introduced to the market within the past five years, may be the worst of your available options:

A. **Native Hadoop BI software is less mature than traditional enterprise BI software, and it lacks many of the features users consider essential.**

B. **Unless your organization has never invested in BI tools, you introduce complexity:**

- End users must switch tools to work with data in Hadoop.
- To your executive team, it's YABIT: Yet Another BI Tool.
- It will take considerable time to drive value.
- Just selecting a new tool will take months.
- On average, it takes 73 business days to go live with a new BI tool. Since Native Hadoop BI platforms are less mature than leading tools, your experience could be much worse.
- The result of your efforts will be another BI silo. That's not smart.

Following the conventional wisdom will not deliver the results your executive team expects. Could there be another way?

Guiding Principles: BI on Big Data

Feeling desperate yet? Don't fret! there is an alternative path: BI on Big Data.

Instead of forcing your business users to conform their BI practices to your data platform strategy, your BI strategy should empower business users to work live with any data, of any size and type, in any data platform, at top speed, and under enterprise governance standards.

The guiding principles for BI on Big Data are:

- **Decouple end user tools and data platforms.**
- **Do not move data.**
- **Push BI workloads into the data platform.**
- **Where possible, leverage open source query engines.**
- **Support BI across modern Big Data platforms.**
- **Preserve existing BI capabilities.**

Let's look at each of these in turn.

Decouple end user tools and data platforms

Enterprise data environments are complex and diverse, with many different data platforms: relational databases, data warehouse appliances, Hadoop, NoSQL databases, in-memory databases and cloud data stores. Moreover, those platforms are constantly changing as CIOs take steps to manage data efficiently. Business users, on the other hand, need stability; it's impractical to make them use a different BI tool for each data platform, or frequently change BI tools. Decoupling is the answer: avoid architectures that create BI silos.

Do not move data

There is so much data today that architectures that make you move or copy data are DOA – dead on arrival. Data movement adds cost, complexity, and latency to a BI architecture, and your goal should be to eliminate it as much as possible. Your users should be able to work with data wherever it is without moving the source data itself.

Push BI workloads into the data platform

Your data platform must support BI workloads; any other computational approach will force you to move or copy data from one platform to another. Data platform developers work tirelessly to develop high-performance query engines.

OPEN SOURCE

Leverage open source query engines

As much as possible, your BI architecture should use open-source query engines, such as Hive on Tez, Spark, Impala or Presto. The cadence of performance improvement for these engines is remarkable; for example, Spark SQL performance doubled from Spark 1.6 to Spark 2.0. Other open source query engines show similar gains. Curious to see what the latest research shows? Get a copy of the test benchmarks here.

Support BI across modern Big Data platforms

Your Big Data strategy is complicated and manages many different types of data across platforms, on-premises and in the cloud. Your approach to BI should enable your business users to answer questions regardless of where you choose to manage data.

YABIT
(Yet Another BI Tool)

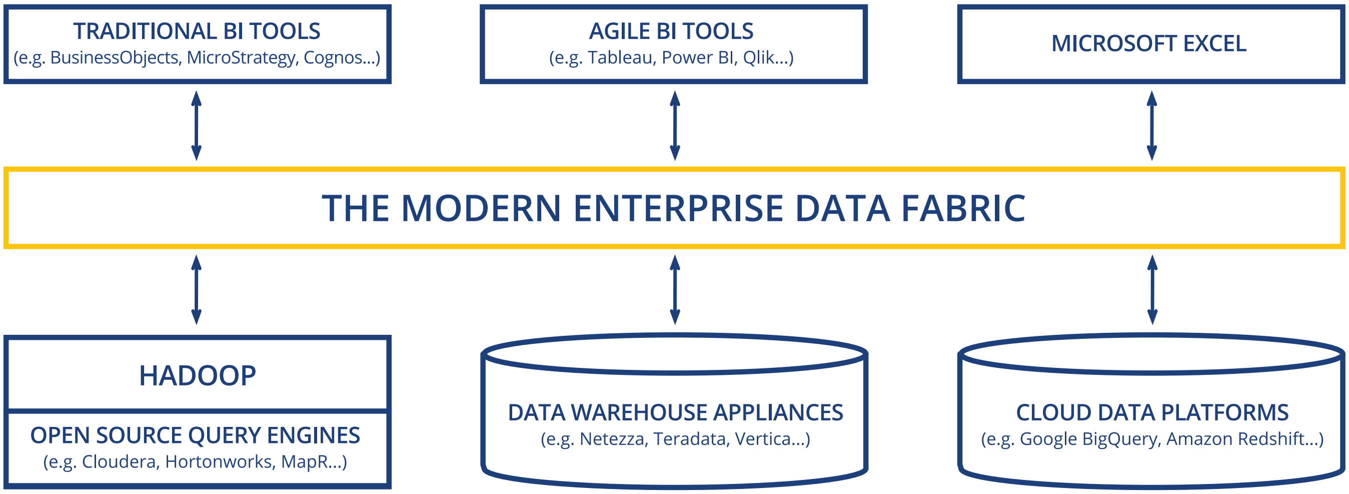
Preserve existing BI capabilities. At all costs, avoid YABIT.

Your organization has invested time and money getting business users up to speed with tools like Tableau, Microsoft Excel, Qlik, Spotfire, MicroStrategy, PowerBI, JasperSoft, SAP Business Objects and IBM Cognos. Introducing another siloed tool into the mix is a mistake.

The core of your approach to BI-across-Big Data is a Modern Enterprise Data Fabric

A Modern Enterprise Data Fabric is a virtual layer that interacts with BI software and data platforms, brokering queries between end users and data sources. It unifies, retains and manages all the critical business concepts, such as hierarchies, organization charts, fiscal calendars and KPIs, so they are accessible to all users. It also handles security across platforms, and serves as a cache for optimal performance. Figure 1, below, illustrates this concept.

Figure 1: The Business Semantic Layer


TRADITIONAL BI TOOLS
(e.g. BusinessObjects, MicroStrategy, Cognos...)
AGILE BI TOOLS
(e.g. Tableau, Power BI, Qlik...)
MICROSOFT EXCEL
THE MODERN ENTERPRISE DATA FABRIC
HADOOP
OPEN SOURCE QUERY ENGINES
(e.g. Cloudera, Hortonworks, MapR...)
DATA WAREHOUSE APPLIANCES
(e.g. Netezza, Teradata, Vertica...)
CLOUD DATA PLATFORMS
(e.g. Google BigQuery, Amazon Redshift...)

An approach to BI based on these four principles empowers your business users. It radically reduces time to value by eliminating effort in four areas.

- Evaluating and selecting new BI software
- Building new data feeds
- Training, mentoring and coaching end users
- Rebuilding existing BI assets from scratch

This approach is the only one that delivers the power of BI from your Hadoop platform – within the window of time your executive team expects.

And Now What?!

Change Doesn't Occur without Action. Use the below resources to guide your journey...or better yet, apply to get a complimentary "Big Data Maturity" Assessment scheduled for your organization here.

How Enterprises Make BI on Big Data Work

How Yellow Pages Does BI on Hadoop

How Macy's Innovated for Scale, Speed & Simplicity

How Quotient's Big Data Leader Drove a BI on Hadoop Evolution

Get technical: Architect, Scale and Turbo-Charge BI on Big Data

The Ideal Architecture for BI on Big Data

5 Ways To Scale BI on Hadoop

Turbo-Charge BI on Hadoop

No Time? Get the Quick Download via the "BI on Big Data Essentials Videos"

- Don't Move Data: de-layering your stack and the value of one semantic layer
- Scale Out, Not Up: using aggregates: the speed of OLAP, the scale of Hadoop
- Schema On Demand: agility requirements and working with complex data types on the fly
- Stay Open: future-proof your stack: work with all flavors of Hadoop, all flavors of BI

About the Author

Thomas W. Dinsmore is an independent consultant serving startups and investors in the advanced analytics market. Before launching his consultancy, Mr. Dinsmore served clients for The Boston Consulting Group, PricewaterhouseCoopers, Oliver Wyman, IBM Big Data Solutions and SAS Professional Services.

Thomas has led or contributed to analytic solutions for clients across vertical markets, including AT&T, Banco Santander, Citibank, Dell, J.C.Penney, Monsanto, Morgan Stanley, Office Depot, Sony, Staples, United Health Group, UBS, Vodafone and many other clients around the world.

Mr. Dinsmore's book, Disruptive Analytics, recently published by Apress, is available now on Amazon.