Business white paper

Five Requirements For Big Data Platforms

Harness the power of 100 percent of your data to support new challenges

Big data is both a challenge and an opportunity. On the challenge side, the onslaught of data can overwhelm your organization if you don't have systems in place to automatically capture, store, manage, and analyze all of it. On the opportunity side, big data can be the key to competitive advantage—if you have the right big data platform in place.

Table of contents

Executive summary

Organizations of all sizes face a common set of challenges brought by an onslaught of big data. Architectures are outdated, the flood of data is increasingly diverse, and technology choices have proliferated, resulting in a growing gap between human information and technology. However, organizations know that by overcoming the challenges, the potential to transform and modernize the business is immense. Big data contains the building blocks for businesses to drive new revenue, increase efficiency, and better comply with federal, state, and industry regulations.

Most current vendor approaches fall short of providing purpose-built big data platforms. Legacy technology simply can't keep pace with the volume, velocity, and variety of big data, and many newer technologies come with all the trappings of emerging and immature technologies. What's needed is a new approach to the big data challenge. That's the point of this white paper. It outlines the essential requirements for a platform that allows you to harness the power of 100 percent of your data. This document contains a manifesto of sorts—a summation of the HP big data platform philosophy that guides our product direction.

There are five fundamental requirements for a big data platform built for the realities of today and beyond. Specifically, a big data platform should enable your organization to:

1. Fully harness 100 percent of your data
2. Seamlessly develop, deploy, and consume data anywhere
3. Achieve scale and speed without compromise
4. Leverage open, extensible, developer-oriented technology
5. Realize economics that work

The big data problem

Today's businesses face a variety of unprecedented challenges in putting big data to use. Often, just defining the challenges can be difficult, particularly when they keep changing. Yet understanding and adapting to these challenges are essential steps in the process of selecting a big data platform.

The challenges of big data run beyond the simple fact that there is more data being created than ever before. Being ready for big data also means dealing with new varieties of information from web logs, call center voice data, video, Internet-connected devices, and more. It also speaks to the unpredictability of data. How much data will you have to deal with tomorrow, next month, or five years from now? Will you be able to scale to meet the needs of your organization and your customers?

So what about those challenges? Here is a brief overview of why businesses globally are struggling to characterize big data:

Outdated architectures
In their efforts to turn mountains of data into actionable business insights, enterprises are held back by legacy architectures that weren't designed for the volume, velocity, and variety of the era of big data. Technology architecture tends to lag behind both business needs and vendor innovation. As a result, many organizations struggle to match their architectures with their evolving needs.

Increasingly diverse data
Over the past 10 to 15 years, the world has changed in fundamental ways. While the volume of data available today is significantly larger than it was even a decade ago, today's data is untamed. It comes in all forms, conforms to a plethora of standards, and comes at us from all directions—from social media and the Internet of things to enterprise business systems and end-user applications.

Data bound to applications
Legacy applications historically chain the data to the application, resulting in a closed "single source of truth," which is anything but. These applications are rigid, difficult to change, and difficult to integrate with modern apps and services, and so is their associated data. Organizations typically have many of these, resulting in silos of information that can't be put together in meaningful ways. Significant funds are spent on master data management systems that are still limited to traditional business data and can't provide comprehensive governance of all data across structured and unstructured data, which is often a prerequisite to figuring out which of this legacy data is really worthwhile.

A proliferation of technology choices
Over the past decade, the rate of technology innovation has increased significantly, leading to dramatic growth in the number of choices for capturing, preparing, and analyzing data. The solutions have unique use cases, system requirements, and cost models. Companies may wonder whether to try to enhance legacy technologies to handle bigger data, but may also be dealing with high licensing costs. Instead, many consider adopting open source technologies like Hadoop, with its low licensing costs and easy hardware requirements but arguably higher costs to hire experienced Hadoop professionals. They may be unaware of other, more suitable solutions that can help meet the challenge of big data.

New business needs
Digital business opens up opportunities to take the data-driven enterprise to a whole new level. What was manual is now automated, what is automated can be instrumented, and what is instrumented can be analyzed. These advances create a closed feedback loop in many industries that simply didn't exist before—from evidence-based medicine to online games that change in real time. Organizations are now challenged to find ways to take advantage of these data-driven opportunities.

As a result of these trends, there is a gap between human information and technology. The growth rate in data generation is such that we now double the global amount of information just about every three years. To complicate the picture, unstructured data (human information, in diverse forms, from voice and video files to knowledge worker documents) and semi- and poly-structured data (such as machine data and complex medical records) is now 10 times larger than structured business data and is growing 10 times as quickly.

Why current approaches fall short

Rather than making things simpler, the recent proliferation of technology choices has only made things more complex. Having defined the problems of big data in the previous section, here we will clarify why current technologies fall short of being true big data platforms:

Legacy technology was almost always purpose-built to be a silo. Often referred to as the "single source of truth," these silos were designed to be narrowly defined repositories for things like financial reporting or supply-chain management or customer detailed records, aligning with organizational boundaries. Each repository would have its own data steward, data formats, and compliance policies. Essentially, each silo lived in its own separate world.

These technologies were frequently built on proprietary architectures, and sometimes even on specialized hardware, making integration costly and dissuading all but the most well-resourced and skilled. As a result, they can't scale up or across in the way organizations now require. Furthermore, they were built in a time when almost all the information stored was highly structured—such as the data housed in the enterprise resource planning (ERP) and customer relationship management (CRM) systems used widely in enterprise environments. They can't accommodate the deluge of unstructured and semi-structured data coming in from every direction.

Without question, legacy techniques for data management are falling short, as are decisions made solely on their recommendations. Old technologies simply can't handle the requirements of today's big data challenge—especially when it comes to machine logs and human information.

On the other hand, emerging technologies today are largely focused on acting as low-cost disruptors in the area of data storage, and they often lack in analytic richness and performance. They tend to be built on open technology stacks, which can lower upfront acquisition costs, and they implement bare bones mechanisms for distributed data processing. However, they often suffer from a lack of unified design, which can drive up the costs of adoption as users struggle to get value from them. So while they may scale well in terms of data storage, and provide favorable economics for this, they do not provide the full range of capabilities required by a big data platform.

To move forward, businesses need to move out of the age of legacy technologies, sensibly leverage niche emerging technologies, and enter the era of the enterprise big data analytics platform.

Five imperatives for your big data platform

1. Fully harness 100 percent of your data

The rate of data growth continues to accelerate, as does its complexity. As organizations become more competent about extracting insights from their conventional data—such as structured data from the enterprise data warehouse, machine logs, and web logs—they rapidly encounter diminishing returns. While they can quickly deliver powerful insights when they develop clickstream analytics for the first time, finding additional insights of equal power quickly becomes more difficult. To continue their progress, organizations are looking further afield toward human data historically considered too difficult to analyze, such as video, audio, and text—typically referred to as "unstructured" data.

Today, technology exists to create structured data from unstructured data—a key form of enrichment that allows the data to be linked with other structured data. Once this is done, entirely new insights are possible. Consider a model predicting churn. A decade ago such a model might be based on purchase history. Two years ago, it might incorporate website purchase activity and browsing behavior. Today, it might incorporate audio from customer interaction with a call center to identify conversation topics and attitude about the topics—in near real time. In this way a business can develop finely tuned, targeted models to predict churn, and even associate that with individual customer profitability to make a real-time choice about when and how to retain that customer, and whether it's even profitable to do so.

And perhaps, most importantly, obtaining data that can be used for big data analytics is just the starting point. A true analytics platform must include capabilities for preparation, enrichment, and analysis of data—both built-in capabilities as well as extensions, and the ability to extend the platform even further.

2. Seamlessly develop, deploy, and consume anywhere

Putting data to work effectively has always been highly dependent on the ease of use of the analytic platform. Exotic development languages or niche platforms may lend themselves to solving one or two use cases quickly, but these don't give you the ability to create true data transparency—in which developers, data scientists, analysts, leaders, and other users do not need to know where the data is located or how it is stored.

This transparency greatly simplifies the development and use of complex analytics, which means you can deliver analytics with higher quality, in less time, and at a lower overall cost. Furthermore, this transparency aligns with a shift many businesses are pursuing to make analytics a core competency, which requires that insights be truly ubiquitous. It is only in this way that a business can effectively base decision-making on analytics.

Other essentials for seamless cloud services include flexible deployment models—such as on-premises, hybrid cloud, and cloud solutions with pre-configured hardware and software. This gives you the option to use one model or another based on your business requirements. For example, some organizations start out using external cloud services and then bring the initiative back in house. A true big data platform supports this flexibility, while preserving all customizations. This is quite unlike today's services that have no option beyond cloud.

3. Achieve scale and speed without compromise

As much as vendors like to talk about higher value topics—like push-button "automagic" analytics and pretty graphics—one of the primary attributes of big data is that it is … well, big. So as important as appealing dashboards or pre-built analytics may be, if the technology can't handle the scale of the new world of data and deliver exceptional performance it's a limiter, not an enabler.

Organizations that are successfully on the path to big data analytics don't compromise between scale, speed, and analytic richness. Rather, they insist on having all these capabilities in the same platform. They load terabytes per hour to petabyte-scale repositories, deliver insights within minutes of the data being created in the source systems, allow thousands of users real-time access to the data, and so on.

Looking ahead, the rate of growth in big data is only increasing. A decade ago, a terabyte of data was big. Today, a petabyte is big. And very quickly, organizations worldwide will surpass petabyte scale. So it's important not just to deliver performance and scale today, but to deliver a platform that can grow and evolve as big data grows and evolves. In this way, as the market moves past petabyte-scale repositories, your organization can rely on your technology scaling with you.

Analytics is not just about asking a single question and getting a single answer; it's really about following iterations of questions and answers to new innovations forged on new paths, previously unknown due to the glacial speed at which analytics was performed across incomplete data sets.

Other key requirements for achieving scale and speed without compromise include:

- The ability to load and query simultaneously, so that systems do not need to be taken offline
- Data compression capabilities that not only allow more data to be stored per node, but reduce the impact of a server's biggest bottleneck—I/O
- The ability to run on a variety of hardware, which gives your organization more choices and more ways to make the economics work

It's worth noting that achieving scale and speed should be possible without driving up your total cost of ownership. For example, enterprise data warehouses can increase their speed by adding more specialized hardware, but doing so results in higher TCO and leaves you with just a stopgap measure.

4. Leverage open, extensible, developer-oriented technology
Big data represents a potential architectural change for many organizations. If your organization is going to pull together data never before analyzed and make it available to everyone, there are some qualities the system requires to meet your business needs.

For starters, a big data platform must be open—so it can operate with a diverse array of other tools and technologies. In most businesses there are existing systems, such as databases and business intelligence systems, that aren't going to go away, so it's important to work with these. Similarly, the platform should support or integrate with open source tools, such as R and Hadoop, and a choice of data visualization and ETL tools.

In another important requirement, the platform must be extensible. Just as much as the world of big data is changing quickly, bringing a constant stream of new demands, so is business. This means the platform you deploy today will very likely need to do something differently tomorrow. An extensible system makes it simpler for the vendor to deliver new functionality, and it also allows your organization to do so on your own when you have requirements particular to your business.

To us at HP, big data is only as useful as the answers it can provide. To this end, putting developers first in product plans is a key requirement of a true big data platform. One way to do that is to ensure that the platform offers open APIs, community support, and a marketplace with extensions and integrations. A true big data platform must allow users to interact with data, even big data, in a conversational manner and explore both unstructured and structured new data sources quickly and easy. This capability coupled with visualization tools enables powerful visualization and interaction with data to answer questions that no one has even thought to ask yet. "Conversations with data" will revolutionize your approach to big data and big data analytics.

Finally, technology that sits on the shelf—or "shelfware"—does not do a business any good. To ensure that you can gain maximum value from your investment, your big data platform should be simple to adopt, should fit with governance models, should provide standard mechanisms for security and workload management, and should provide flexible deployment options, such as on-premises, cloud, SaaS, or hosted.

5. Realize economics that work

One of the more significant challenges organizations face in mastering big data is the economics of the platform. Many businesses built enterprise data warehouses or other repositories in the last 20 years, using technology that was state of the art at the time. These legacy vendors often created a complicated pricing model that charges extra for data volumes, CPUs, users, connected systems, connected applications, and more. It's a pricing model that also charges a percentage of the original purchase in license and maintenance year after year. As a result, the cost per byte with these legacy platforms is simply too high and grows higher as you try to handle big data.

It's not simply a matter of acquiring lower-cost, open-standard hardware—though this can be a good thing. And it's not simply a matter of shifting to a new architecture that puts more of the data closer to the processor. These are simply Band-Aids on aging technologies from largely the same vendors promoting these outdated platforms.

What you really need is a big data platform and ecosystem that can deliver full functionality with an economic footprint that makes sense for the use case. It's not just about a single technology but a purposefully designed ecosystem to meet diverse needs, such as extremely low-cost simple retention alongside state-of-the-art analytics. And the technologies themselves must be designed with a target economic model in mind so the vendor can provide sustained support, development, and innovation that suits marketplace needs. What's more, the big data platform should be standards-based, so there's limited to no retraining needed, and should be easy to administer, so the database admins (DBAs) and others who use the system can focus on more challenging and beneficial duties—rather than running batch reports overnight.

Finally, the technology is only part of the economic picture. It needs to be implemented, maintained, and used effectively. A big data platform sufficiently esoteric that only a hundred people globally know how to implement it may be free of license cost, and yet be highly expensive to use. It is therefore important to look beyond the platform at the vendor's capabilities to help design and deliver the technology.

HP Haven: a platform built for the era of big data

The HP Haven Big Data platform is designed to meet all of these requirements to be a true big data platform built for the era of big data. HP Haven is designed to harness 100 percent of your data—business, human, and machine—to inform every decision and help you capitalize on opportunities and solve the toughest challenges. Available on-premises or in the cloud, HP Haven offers big data analytics and next-generation applications at unmatched speed and scale.

HP Haven is the industry's first comprehensive, scalable, open, and secure platform for dig data. The platform comprises software, services, and hardware components that work together to enable you to analyze all of the data that is relevant to your organization, regardless of its source. It can analyze internal business data, machine data/sensor data, and unstructured human information, such as social media sentiment, email, video, and voice recordings. With hundreds of connectors, HP Haven is designed to easily ingest data from diverse sources.

The capabilities of the HP Haven Big Data platform are available in the HP Haven Enterprise offering and the HP Haven OnDemand suite of cloud services.

HP Haven Enterprise
Powered by industry-leading analytics engines, the Haven platform enables you to:

- Connect quickly to more than 700 source systems with out-of-the-box connectors, streamlining and accelerating data integration.
- Index, analyze, correlate, and add context to your data to make sense of it all.
- Load and query your data simultaneously to gain speed, scale, and cost advantages, or index it in place.
- Develop applications on your data, using a set of well-defined, open, programming interfaces and open-source standards such as Hadoop.

HP Haven OnDemand
With HP Haven OnDemand, you can tap into key components of the HP Haven Enterprise big data platform to gain blazing fast insights, rapid time to value, and analytic functionality on all types of data, within minutes. HP Haven OnDemand offerings include:

- HP Vertica OnDemand, which delivers enterprise-class big data analytics via the cloud, including an extensive set of built-in analytic capabilities with no compromises on performance and flexibility
- HP IDOL OnDemand, which provides the industry's most comprehensive set of big data web services that you can use to build next-generation applications that can analyze a broad spectrum of data types

HP Haven on Hadoop
Leveraging years of experience in big data analytics, this offering taps into the full power of Hadoop. By offering a rich, fast, and enterprise-ready implementation of analytics on Hadoop, you can perform analytics regardless of where the data is stored, the format, or Hadoop distribution.

Running on the HP Haven platform means that you can deploy fast in the cloud or host your own solution within your own four walls. It also offers the flexibility to leverage your Hadoop investments for more robust analytics.

An assemblage of powerful analytics engines

At the core of the HP Haven Big Data platform are three analytics engines working in harmony to solve big data challenges. You can leverage one, two, or all three of them, depending on your requirements:

- HP IDOL powers analytics, digital marketing, information management, and governance solutions by enabling you to index, search, and analyze human information at scale and in context. It can process hundreds of file types, including tweets, email, audio, images, and video.
- HP Vertica is a massively scalable MPP database, custom-built for real-time analytics on petabyte-sized data sets. It supports standard SQL and advanced analytics, and it offers support for all leading BI and ETL vendors.
- HP Distributed R is a framework to parallelize machine learning algorithms frequently used with the R statistical analysis platform. It interacts seamlessly with R, and includes a high-speed connector for moving data to and from the HP Vertica platform.

Key takeaways

While today's data deluge brings unprecedented challenges, it simultaneously creates rich opportunities for organizations that approach the problem strategically and put the right big data platform in place.

That platform should enable your organization to:

1. Fully harness 100 percent of your data
2. Seamlessly develop, deploy, and consume data anywhere
3. Achieve scale and speed without compromise
4. Leverage open, extensible technology that is developer friendly
5. Realize economics that work

As the industry's first comprehensive, scalable, open, and secure platform for dig data, HP Haven is designed to meet all of these requirements for the era of big data.

Learn more at
hp.com/haven

Sign up for updates
hp.com/go/getupdated

Rate this document

4AA5-8476ENW, June 2015, Rev. 1