insideBIGDATA

*InsideBIGDATA Guide to*

# The Intelligent Use of Big Data on an Industrial Scale

*Written by Peter ffoulkes*

BROUGHT TO YOU BY

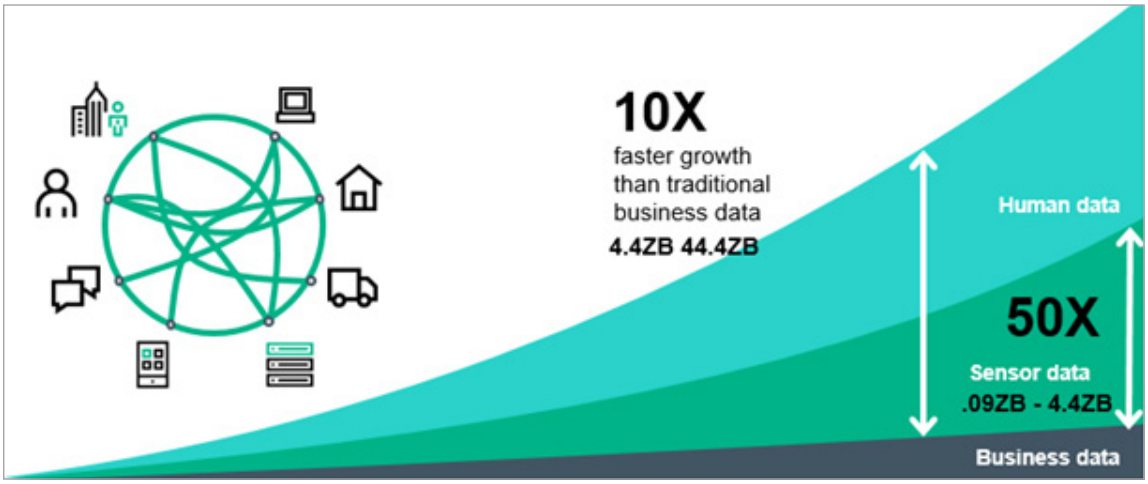
Hewlett Packard Enterprise

# The Intelligent Use of Big Data on an Industrial Scale

Several decades ago, Saudi Oil Minister Sheikh Yamani gained recognition for his insight into global development: "The Stone Age did not end for lack of stone, and the Oil Age will end long before the world runs out of oil." Today, we live in what many call the Information Age, and we are in absolutely no danger of running out of information, particularly in data form. There is a general perception that we are overwhelmed with data, making the ability to store, process, analyze, interpret, consume, and act upon that data a primary concern. For large-scale, multi-national organizations and those in heavily regulated industries — such as finance, healthcare, or those covering multiple industry verticals — the situation becomes even more complex and challenging. Escalating data concerns are rampant in the Internet of Things (IoT) Age, during which increased quantities of available data are exceeding the capacity of traditional computing. The question then becomes, how do we consume those data sources and transform them into actionable information?

## The Exponential Growth of Data

There are many sources that predict exponential data growth toward 2020 and beyond. Yet they are all in broad agreement that the size of the digital universe will double every two years at least, a 50-fold growth from 2010 to 2020. Human- and machine-generated data is experiencing an overall 10x faster growth rate than traditional business data, and machine data is increasing even more rapidly at 50x the growth rate.




## Contents





The acquisition and analysis of data and its subsequent transformation into actionable insight is a complex workflow which extends beyond data centers, to the edge, and into the cloud in a seamless hybrid environment. The utilization of edge devices, in situ-computation and analysis, centralized storage and analysis, and deep learning methodologies which accelerate data processing at scale requires a new technological approach. Historically, data processing and analytics systems had specialized features for business analytics and high-performance computing (HPC) workloads. Yet with the advent of big data and industry-standard x86-based computing, we are seeing a convergence in big compute, big data, and IoT for analytics. IDC research categorizes this convergence as high-performance data analytics (HPDA).

The key factor driving the adoption of data-intensive computing is the need to rapidly analyze exploding volumes of data at the point of creation and at scale.

The HPDA market is at the center of big data analytics. The key factor driving the adoption of data-intensive computing is the need to rapidly analyze exploding volumes of data at the point of creation and at scale. An important consequence of this explosion is the need for users to adopt advanced data analytics technologies. Enterprises now have access to cheaper and more powerful computing platforms, and modern analytics software like Hadoop and Spark enable real-time analytics for a wide range of use cases, including fraud and anomaly detection, business intelligence, affinity marketing, product design and development, process automation, and personalized medicine. In addition to these software frameworks, implementing storage capacities and capabilities which enhance data flow, in-place analytics, and storage efficiency such as object storage and high-performance distributed file systems, is critical for effective scaling.

According to IDC's survey on the most important digital transformation projects, respondents cited cloud transformation/transition (66%), IoT (32%), and big data/cognitive solutions (27%) as key initiatives for big data usage and development. The cloud provides scalability, and the IoT forms the foundation for investments in big data and cognitive computing. IDC predicts that by 2020 50% of all business analytics software will incorporate prescriptive analytics built on cognitive computing technology, and the amount of high-value data will double, making 60% of information delivered to decision makers actionable.

## Data Growth Challenges

The exploding volume and speed of data growth has introduced several challenges:

- System management and growing cluster complexity
- Data center power, cooling, and floor space limitations
- Storage, data movement, and management complexity
- Lack of support for heterogeneous environment and accelerators
- Significant shortage of skills to integrate and manage the big data ecosystem

## Infrastructure Drives Improvements

Organizations are evaluating and implementing infrastructure to drive the following improvements:

- Manage growth and operational cost of big data infrastructure
- Provide elasticity and flexible capacity
- Ensure performance for diverse workloads
- Rapidly deploy and scale infrastructure
- Simplify management with Big Data as a Service (BDaaS)

In this document, our focus is on "industrializing" big data infrastructure — bringing operational maturity to the Hadoop data ecosystem, making it easier and cost-effective to deploy at enterprise scale, and moving companies from the proof of concept (PoC) stage into production-ready deployments.

## The Changing Data Landscape

Large enterprise customers have made huge investments in data warehousing technology over the past decade. The cost of upgrading and adding new data warehouse licensing is cost prohibitive. Technologies like Hadoop and Spark offer organizations the option of accruing a greater ROI on existing data warehouse resources, by providing a data processing platform to offload Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) workloads from the data warehouse and to reduce the cost of storing and accessing less frequently used data.

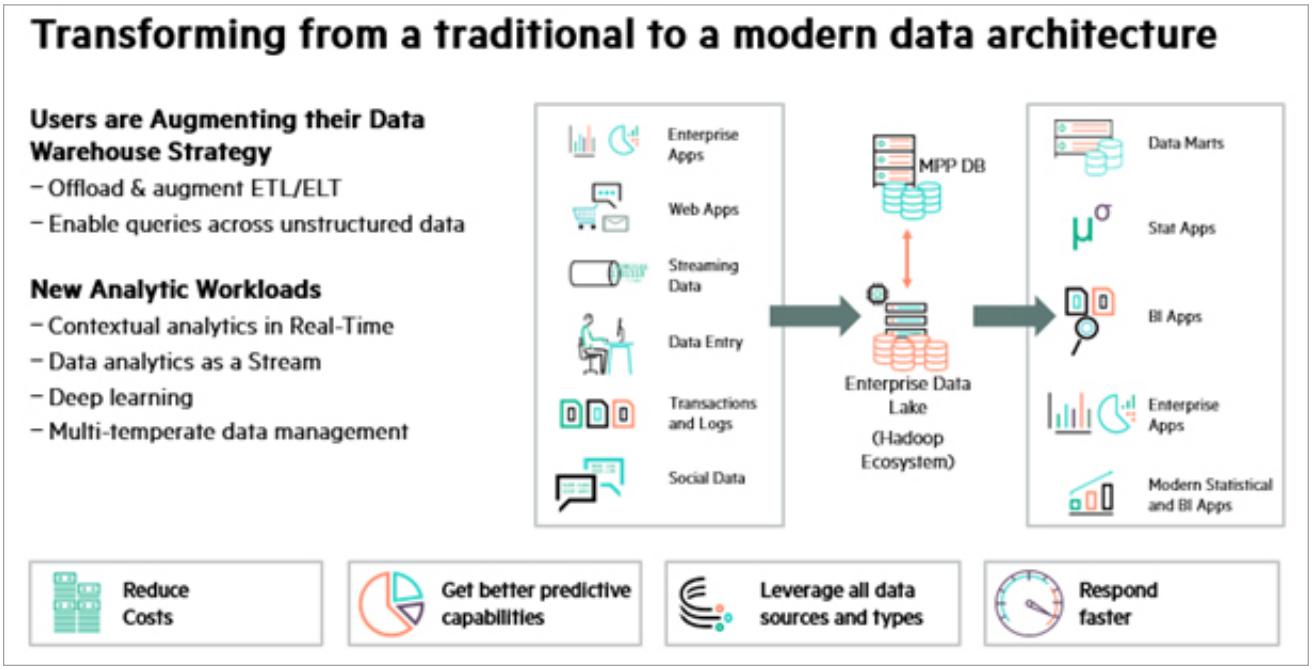
As organizations endeavor to extract insights from a combination of human, machine, and business data, Hadoop, a cost-effective computing and storage platform, becomes a more viable approach for harnessing operational data and big data analytics.

Initially, many organizations implement Hadoop to store data from multiple sources in its original form, with the intent to ELT data for downstream analytics. A large percentage of these developments start as small-scale experiments within isolated business units rather than enterprise initiatives,

As organizations endeavor to extract insights from a combination of human, machine, and business data, Hadoop, a cost-effective computing and storage platform, becomes a more viable approach for harnessing operational data and big data analytics.

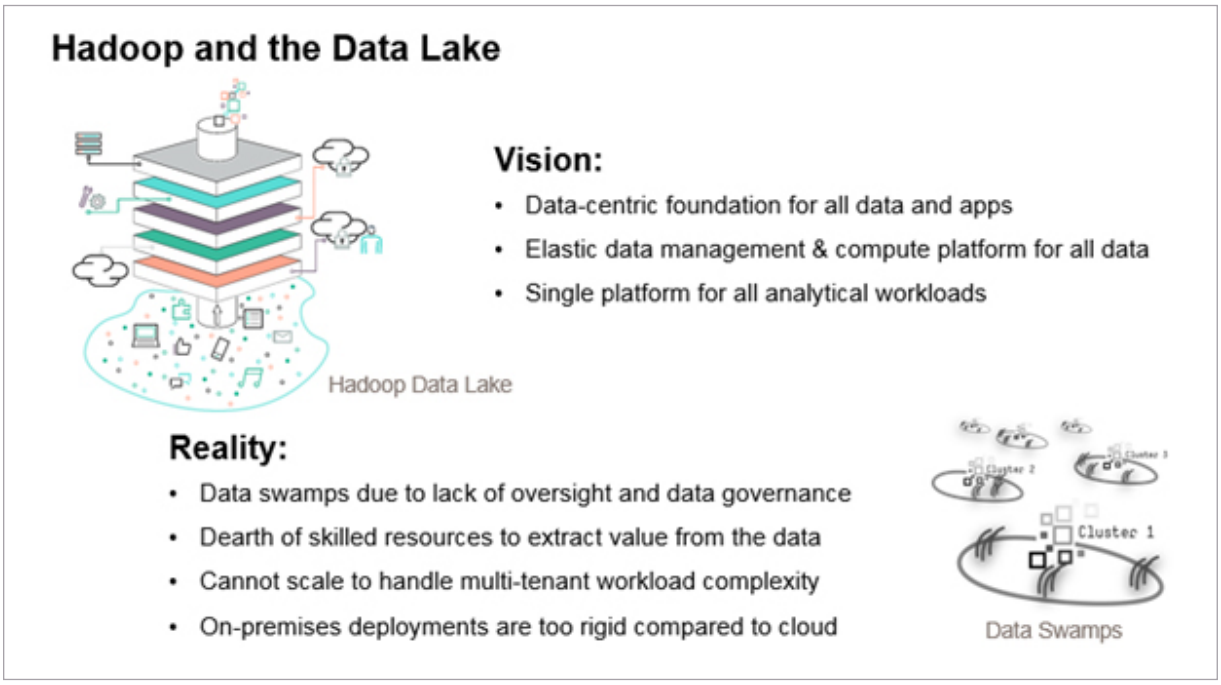
either in the public cloud or as a standalone op-premises deployment to serve a single use case. Data exploration or data warehouse optimization are the most common entry points.

Enterprises with executive sponsorship and a data-driven strategy quickly progress to other analytics use cases that realize greater business value — this includes customer 360, predictive analytics, and data discovery. Data lakes are a logical foundation to drive these analytics use cases, centrally manage a variety of data from multiple sources — both processed and unprocessed (dark data), and enable enterprises to glean insight from the information.








Enterprises have not yet realized the true potential of data lakes for two key reasons:

- Lack of a correct data model that unifies all data
- Traditional Hadoop infrastructure is inefficient and rigid

While vendors promote the ease of storing and analyzing data on shared storage, without the additional step to ingest the data into Hadoop, the reality is that all data required for analytics must be stored and organized in appropriate formats in order to accelerate different analytics workloads.

Conventional platforms are designed for batch workloads and do not scale efficiently for modern analytics workloads (i.e. machine learning, streaming, SQL, and NoSQL interactive analytics). The need for performance and scaling out linearly to accommodate various software frameworks, such as Spark and NoSQL which are more memory- and compute-intensive, requires flexible compute and storage resources capable of consolidating diverse workloads on top of the data lake.

Due to advancements in big data usage and management, many organizations realize that their existing infrastructure is underutilized or overprovisioned, resulting in cluster and data sprawl. The inability to meet the needs of new analytics workloads is driving IT departments towards consolidation. While data lakes consolidate information, cluster and workload consolidation towards a multi-tenant, elastic platform is a more complex undertaking. The advent of IoT and cloud analytics requires additional capabilities that enable more efficient and elastic mechanisms to ingest, store, and process data across remote locations.

The inability to meet the needs of new analytics workloads is driving IT departments towards consolidation. While data lakes consolidate information, cluster and workload consolidation towards a multi-tenant, elastic platform is a more complex undertaking.



## Realizing a Scalable Data Lake

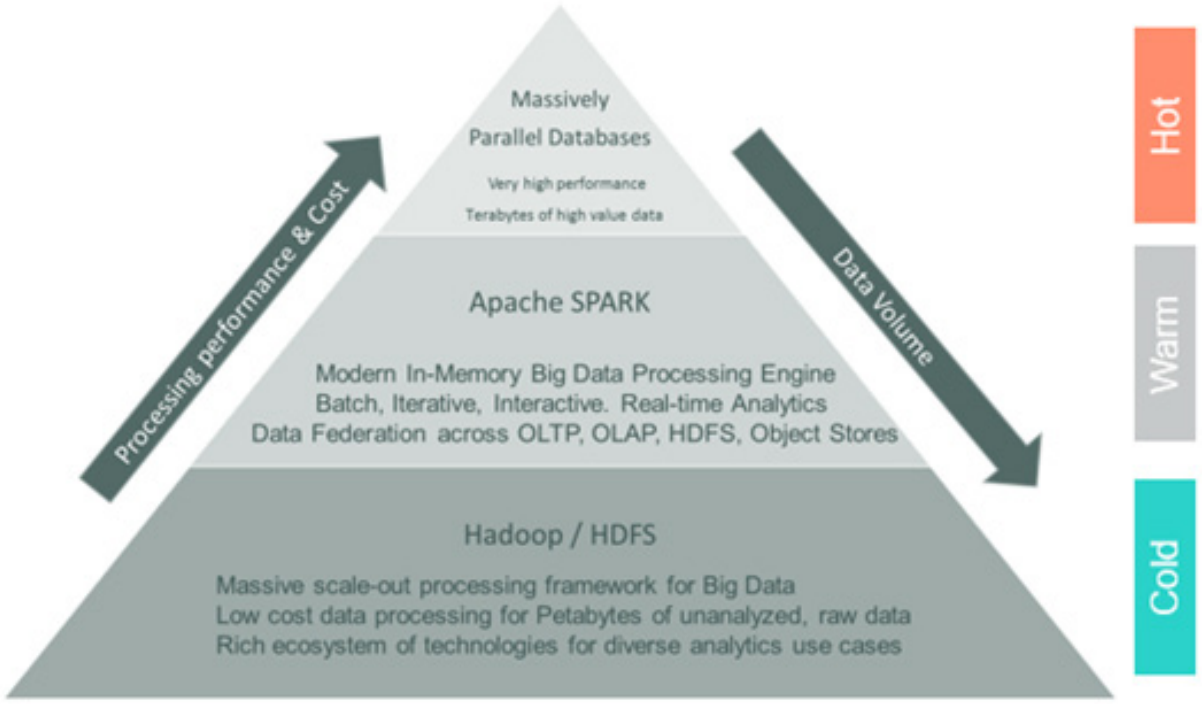
A modern analytics framework must be able to seamlessly manage and analyze data in a number of systems, with the intelligence to selectively move data between data stores based on value, time sensitivity, and cost. Choosing the right technology that optimizes this framework is crucial to delivering real-time access to troves of complex data, with total cost of ownership and ease of management being equally important.

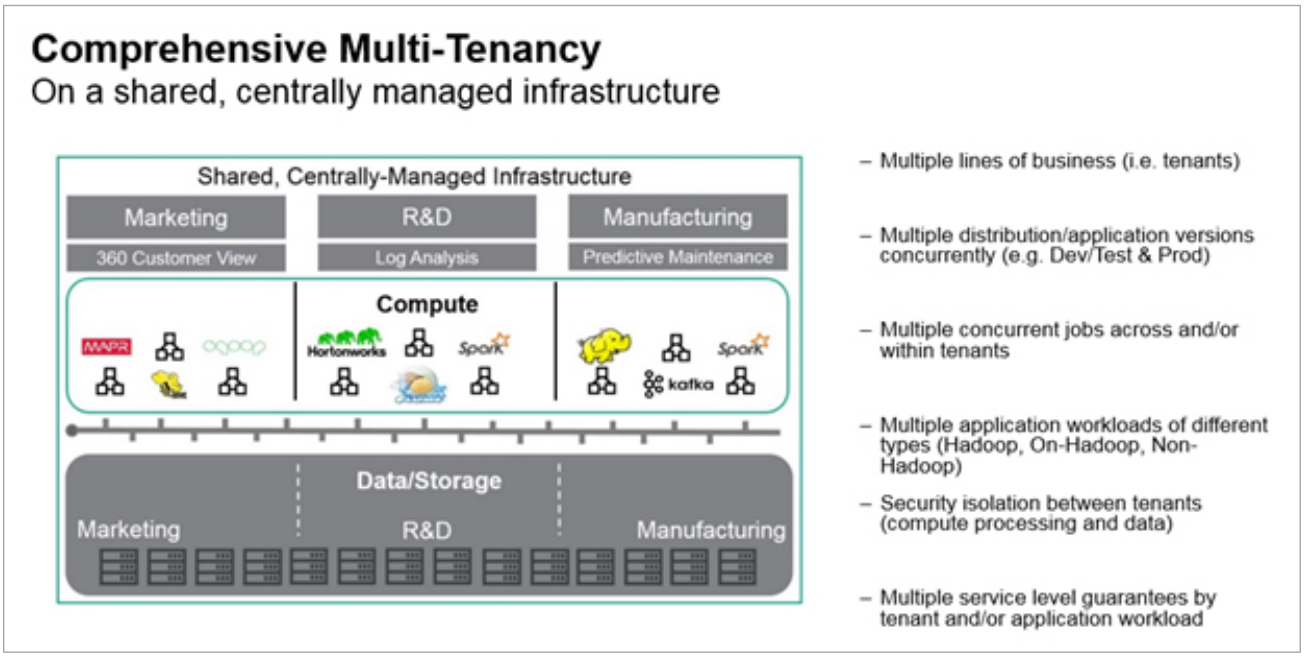
Data can be classified as hot, warm, or cold. Small volumes of business critical data (hot) are processed in massively parallel in-memory databases. Larger volumes of less frequently accessed data (warm) are processed in a low-cost tier, typically flash-based storage and memory. Petabytes of raw interaction data and archived business data (cold) reside in a lower-cost analytics platform like Hadoop, or in an object store.

As the Hadoop ecosystem matures, new processing frameworks like Spark enable a wide range of analytics use cases, from real-time streaming to machine learning, acting as a storage-layer agnostic data federation platform which supports streaming, batch, and iterative analytics.

Since being founded in 2006, Hadoop has proven to be a cost-effective solution for managing data growth across organizations. Rather than buying expensive "latest and greatest" machines to perform massive data workloads, distributing the processing power across a cluster mitigates the need for expensive supercomputers. Hadoop contributors favor data locality, with co-located compute and storage at the node. Each server adds additional compute and storage capacity, scaling linearly. This is what many consider to be a traditional, symmetric architecture where each server is configured identically.

Many organizations share a vision of having all workloads, analytics, and applications running on a common dataset, or a scalable, multi-tenant platform for all data and analytics workloads. This movement begins with consolidating data, using Hadoop as a repository and for simple workloads like ETL and pre-processing data. The variety in the Hadoop software stack can accelerate a diverse set of analytics use cases beyond just ETL and ELT, driving organizations to rapidly evaluate and deploy these technologies to promote new business capabilities. These entities rely on enterprise-grade performance to run various workloads and consolidate data, with the ability to scale these workloads across a common, flexible infrastructure.

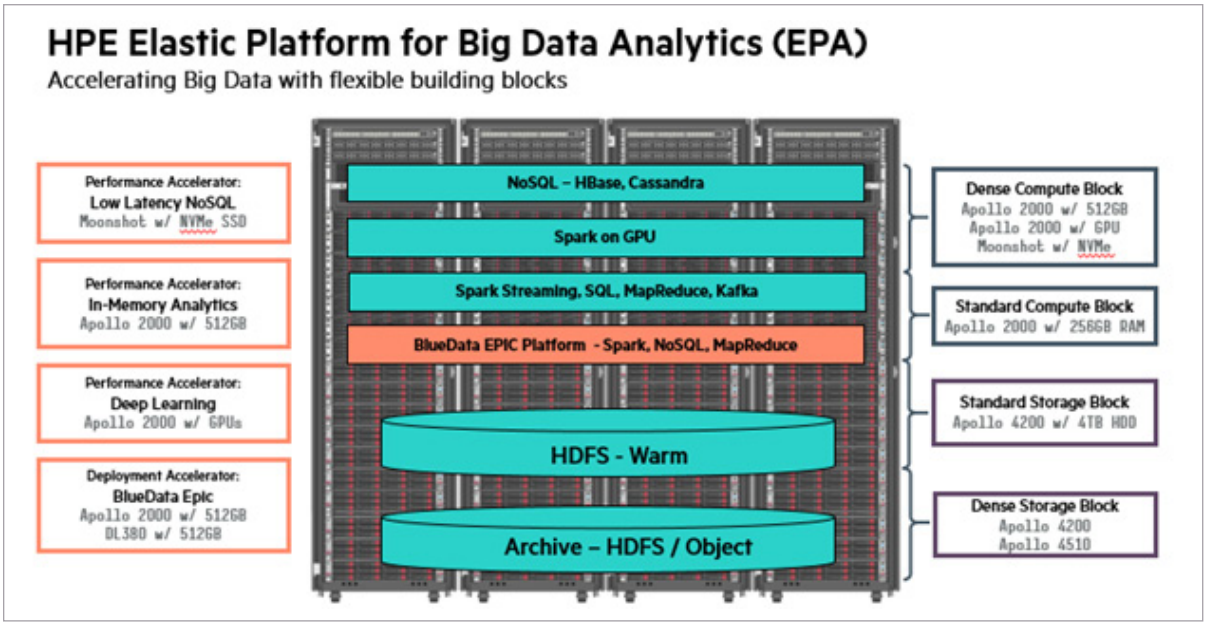





From a platform perspective, it is clear that a homogeneous hardware architecture cannot address all of the required functions:

- Low-latency compute and event processing
- High-latency compute extract-transform-load (ETL) offload and archival storage
- Big memory compute and in-memory data analytics
- HPC compute and deep learning
- HDFS storage
- Archival storage

Organizations are now evaluating hardware options that were traditionally limited to the HPC domain, ranging from general-purpose graphics processing units (GPGPUs) for parallel computing, non-volatile memory express (NVMe), persistent memory for workloads requiring low-latency, and hardware accelerators for offloading compression/de-compression tasks and storage efficiency. This allows for growth and scalability as data volume, variety, and workload needs evolve over time.





## The HPE Elastic Platform for Big Data Analytics

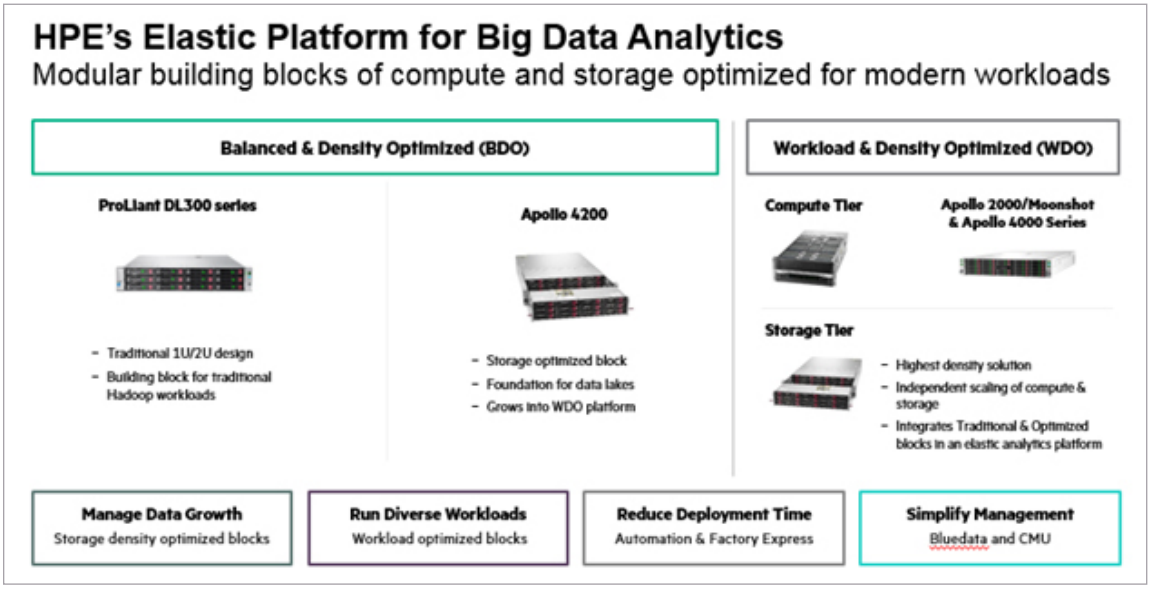
The HPE Elastic Platform for big data analytics is a modular infrastructure foundation that accelerates business insights by enabling organizations to rapidly deploy, efficiently scale, and securely manage the explosive growth of big data workloads.

HPE offers two powerful deployment models under the Elastic Platform:

- **HPE Balanced and Density Optimized (BDO) –** Supports traditional Hadoop deployments that are symmetric (scale compute and storage together), with some flexibility in choice of memory, processor, and storage capacity. This is widely based on the HPE ProLiant DL380 server platform, with density optimized architectures utilizing the HPE Apollo 4000 series servers.
- **HPE Workload and Density Optimized (WDO) –** Optimizes efficiency and price performance through a building block approach. This architecture allows for independent scaling of compute and storage, utilizing the power of faster Ethernet networks while accommodating the independent growth of data and workloads. The standard HPE WDO architecture is based on the HPE Apollo 4200 storage-optimized block and the HPE Apollo 2000 compute-optimized block, coupled through high-speed Ethernet. Combining these linked storage and compute blocks with Hadoop's YARN resource scheduling features delivers a scalable, multi-tenant Hadoop platform. The HPE Apollo 4200 was chosen as the ideal storage block, as it provides exceptional storage density in a 2U form factor. The HPE Apollo 2000 features exceptional compute density, supporting up to four servers with high core-to-memory ratios in a 2U form factor.

Organizations that invest in symmetric configurations have the ability to repurpose existing deployments into a more elastic platform such as the WDO architecture. This system is geared to help customers expand their analytics capabilities by growing compute and/or storage capacity independently, without building a new cluster.

The figure below highlights the various building blocks that create the HPE BDO and WDO system offerings. By leveraging a building block approach, customers can simplify the underlying infrastructure needed to address business initiatives surrounding data warehouse modernization, analytics, and business intelligence, and to build large-scale data lakes with diverse datasets. As workloads and data storage requirements change (often uncorrelated to each other), the HPE WDO system allows customers to add independent compute and storage blocks, which maximizes infrastructure capabilities for data-heavy workloads and promotes seamless scalability.





## Accelerators

Accelerators are an additional component of the HPE Elastic Platform for Analytics. Accelerators are specialized building blocks designed to optimize workload performance, storage efficiency, and deployment. As more demanding workloads are added, accelerator building blocks target intended outcomes. Additional benefits of accelerators include:

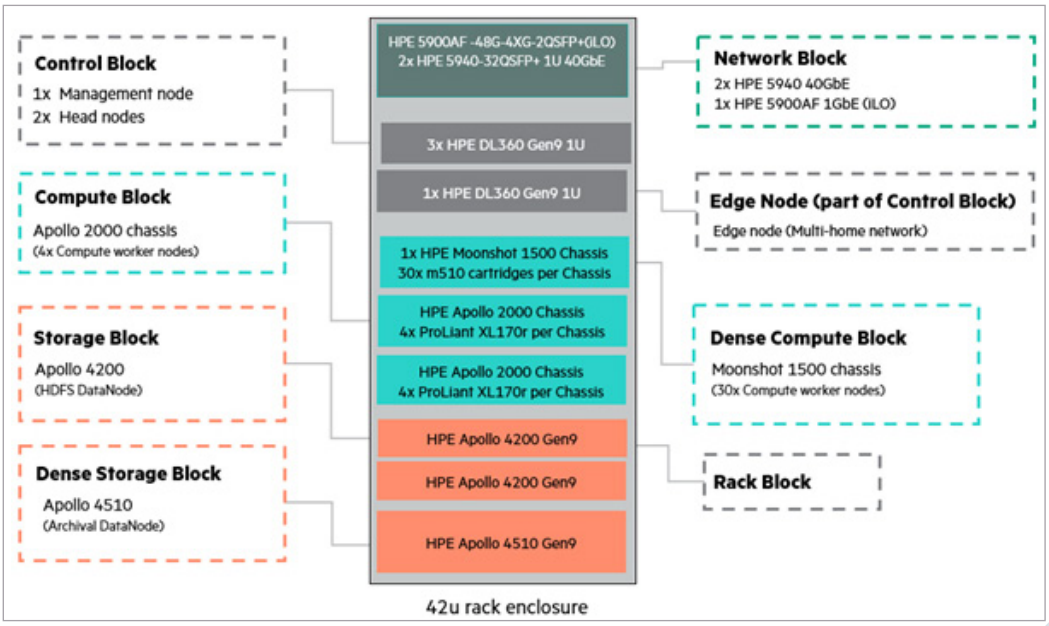
- **Performance acceleration of different workloads** such as NoSQL databases, like HBase or Cassandra, that require low-latency processing in near real-time, in-memory analytics using Spark and/or SAP HANA Vora, deep learning with Spark, and Caffe on GPU accelerated servers
- **Storage efficiency optimization** with HDFS tiering and erasure coding
- **Deployment agility and self-service** through automation and Platform as a Service (PaaS) solutions – HPE Insight CMU and BlueData EPIC respectively

## HPE Workload and Density Optimized System

As Hadoop adoption expands in the enterprise, it is common to see clusters running workloads on a variety of technologies and Hadoop distributions in both development and production environments, leading to issues with data duplication and cluster sprawl. The HPE Workload and Density Optimized system allows for consolidation of data and isolated workload clusters on a shared and centrally managed platform, providing organizations the flexibility to scale compute and storage as required, and using modular building blocks for maximum performance and density. HPE WDO building blocks and accelerators provide the flexibility to optimize each workload and access a central pool of data for batch, interactive, and real-time analytics.

HPE WDO provides maximum elasticity for organizations to build and scale their infrastructure in alignment with the requirements of existing analytics technologies. For example, some pilot environments might deploy a smaller number of symmetric configured storage-density optimized servers for data staging and basic MapReduce workloads. High-latency compute nodes can then be added (repurposing existing symmetric nodes to storage nodes through YARN labels), effectively migrating to an asymmetric architecture, without adding additional storage. A low-latency tier of density optimized accelerator blocks with SSDs or NVMe flash can then be added to address the requirements of time series analysis of large datasets in real-time. By adding accelerator nodes and adjusting compute-to-storage-node ratio by the block, organizations have the ability to tune their cluster toward business initiatives at the rack level. This architecture maximizes modern infrastructure density, agility, efficiency, and performance while minimizing time-to-value, operating costs, and datacenter real estate.

As modern workloads demand evolving levels of storage and compute capacity, the HPE WDO architecture provides density optimized building blocks to target latency, capacity, and performance concerns.





## The Five Blocks of the HPE WDO Solution

The five components of the HPE WDO system include compute blocks, storage blocks, control blocks, network blocks, and rack blocks.

In addition to these standard blocks, HPE has also developed accelerator blocks designed to optimize solution deployment, workload performance, and storage. Unlike the standard compute block (one HPE Apollo 2000 chassis consisting of four XL170r Gen9 servers) and the standard storage block (one HPE Apollo 4200 Gen9, consisting of 28x LFF HDDs or SSDs), optional dense compute, dense storage, or accelerator blocks can be combined to address countless issues — hot/cold storage, high-latency/low-latency compute, NoSQL, deep learning, etc.

Examples of accelerator blocks include the Moonshot 1500 chassis with 30 m510 cartridges for compute, Apollo 2000 with XL170r with 512GB of memory, Apollo 2000 with XL190r with GPUs, Apollo 4200 with 6 or 8TB LFF HDD's, or Apollo 4510 with 3, 4, 6 or 8TB LFF HDD's.

HPE has also developed accelerator blocks designed to optimize solution deployment, workload performance, and storage.

The control block is comprised of three HPE DL360 Gen9 servers, with an optional fourth server acting as an edge or gateway node, depending on the customer enterprise network requirements.

The network block consists of two HPE 5940-32QSFP+ 40Gb switches and one HPE 5900AF-48G-4XG-2QSFP+ 1Gb switch. The system allows for switching these network blocks to 25Gb/100Gb network blocks from HPE or other vendors as long as they have similar configurations.

Finally, the rack block consists of either a 1200mm or 1075mm rack and its accessories.

## Summary

The HPE Elastic platform for big data analytics provides optimal support for modern data processing frameworks including Hadoop, Spark, and NoSQL. A deployment option of this framework, the HPE WDO system enables flexible and independent scale-out of compute and storage, and is ideally suited for deploying and consolidating big data workloads on a multi-tenant analytics platform. And thanks to YARN's multi-tenant capabilities in conjunction with HPE solutions, it is now possible to leverage workload-optimized servers for a variety of use cases, including deep learning with GPU-based or blazing NoSQL performance in a small footprint with the HPE Moonshot.

For existing big data deployments using conventional symmetric architectures, organizations can transition to a truly elastic platform with the HPE WDO System to reduce datacenter footprint, lower operating costs, optimize performance, and efficiently manage rapid data growth. The underlying solution is a common foundation which enables organizations to operate big data infrastructure as a service. For this, HPE provides deployment accelerators with the HPE Insight CMU and the BlueData EPIC software to rapidly provision and deploy an elastic, multi-tenant infrastructure.

The modularity of HPE solutions provides flexibility in block systems that can be manipulated to satisfy workload, density, form-factor, compute, memory, and storage needs in a hybrid environment.

[i] IDC FutureScape: Worldwide Big Data and Analytics 2016 Predictions, Nov 2015, Doc # 259835

