Leveraging big data to gain a competitive advantage across a range of industries

Unlocking Insights with Apache Hadoop

Executive summary

In their efforts to extract value from big data, organizations around the world are turning to the Hadoop data collection, management and analysis platform. Hadoop offers two important services: store any kind of data from any source, inexpensively and at very large scale, and perform sophisticated analysis of that data easily and quickly. This value proposition makes the open source Hadoop ecosystem attractive for diverse use cases across a wide range of industries.

This business-oriented white paper summarizes the wide-ranging benefits of the Hadoop platform, highlights common data processing use cases and explores examples of specific use cases in vertical industries. The information presented here draws on the collective experiences of three leaders in the use of Hadoop technologies—Dell and its partners Cloudera and Intel.

The Hadoop platform

The open source Hadoop platform was originally developed by the world's largest Internet companies to capture and analyze the massive amounts of data they generate. Unlike earlier platforms, Hadoop can store any kind of data in its native format—structured, unstructured or semi-structured—and be used to perform a wide variety of analyses and transformations on that data.

Hadoop allows your organization to store petabytes, and even exabytes, of data cost-effectively. As the amount of data in a cluster grows, you can add new servers with local storage incrementally and inexpensively. Thanks to the use of MapReduce framework, which takes advantage of the parallel processing power of the servers in the cluster, a 100-node Hadoop instance can answer questions on 100 terabytes of data just as quickly as a 10-node instance can answer questions on 10 terabytes.

Both robust and reliable, Hadoop handles hardware and system failures automatically, without losing data or interrupting data analyses. Better still, Hadoop runs on clusters of industry-standard servers. Each has local CPU and storage resources, and each has the flexibility to be configured with the proper balance of CPU, memory, and drive capacity to meet your specific performance needs.

Ultimately, Hadoop makes it possible to conduct the types of analysis that would be impractical or even impossible using virtually any other database or data warehouse. Along the way, Hadoop helps your organization reduce costs and extract more value from your data.

Addressing common data processing challenges

Hadoop is often used to solve two fundamental problems with big data: operational data processing, which addresses common data pain points, and predictive analytics, which answers the big questions. Let's look first at the operational issues.

Hadoop addresses many challenges associated with storing, managing and processing large amounts of data in diverse formats—structured, unstructured and semi-structured. Here's a look at some of the more common operational uses for the Hadoop platform.

Data warehouse optimization

The ETL process can create bottlenecks in enterprise data warehouses. A few heavy jobs can bog down an enterprise data warehouse, and more processing means less query capacity. This processing work can be offloaded to Hadoop to reduce CPU utilization for heavy jobs and to accelerate complex ETL processes. Designed for extreme parallel data processing, Hadoop, when used in conjunction with a tool like Syncsort, can help you speed up ETL processes while reducing costs in comparison to running ETL jobs in a traditional data warehouse.

In addition, Hadoop can serve as an efficient staging and ETL source to complement your existing EDW. Using Hadoop as an enterprise data hub (EDH) to complement to your EDW can drive significant cost savings and other benefits. The goal here isn't to replace your EDW. Rather it is to move certain data, workloads and processes from your existing systems into Hadoop to gain new capabilities and cost economies.

Active archive

With the use of Hadoop as the data-landing zone for ETL offload, Hadoop has become the location of all data in its native format, and therefore a natural archive repository. Hadoop is ideally suited to serve as a single place to store all your data, in any format, at any volume, for as long as you like. While providing cost-effective data archiving, your Hadoop environment can enable broad organizational access to varied data sets for ad-hoc analysis.

Log aggregation

Hadoop is an ideal platform for rapidly growing log aggregation environments. Log aggregation is valuable both when log data is analyzed in real time and used for in-depth analysis. Aggregating and storing logs in a Hadoop environment enables real-time and in-depth analysis, and to an extreme scale. Hadoop excels at high-speed ingestion and is an excellent place to both store and aggregate log data. Both batch and real-time analytics are possible.

Agile data exploration

Hadoop can serve as a staging area that allows users to capture and store new datasets or datasets that have not yet been placed in an enterprise data warehouse. With technologies such as Hive, Impala, Search and Spark, the platform extends the data access funnel to all users in the organization. Users can combine, compose and explore the data to gain new insights.

Use cases in diverse industries

Let's turn now to specific applications of the Hadoop platform for predictive analytics in a wide range of industries. These use cases are based on the experiences of Dell, which has worked on more than 200 Hadoop deployments, Intel and Cloudera, the leader in enterprise analytic data management powered by Apache Hadoop.

The use cases presented here are not intended to be a comprehensive list. Rather, they are examples from a much larger pool of use cases and a much larger range of industries that can benefit from the use of Hadoop in conjunction with sophisticated tools for data analytics, data integration and data management.

Financial services

Unlocking the value of big data

To thrive in uncertain and fluctuating markets, upsell customers, combat fraud, maintain compliance with regulations, and meet other critical goals, financial services firms need to unlock the value of big data. With goals like these in mind, firms are turning to Hadoop to consolidate data traditionally managed at the departmental level to create enterprise data warehouses (EDWs).

Hadoop-based data platforms help financial services firms analyze risk exposure more holistically, comply with regulatory mandates, and perform enterprise-level analytics. The platform provides highly scalable and low-cost data storage and processing that can be tightly integrated with legacy systems.

Hadoop use cases in financial service

- Fraud prevention in credits and payments
- Risk modeling in investment banking
- Cross-selling and upselling in retail banking
- Insurance policy personalization
- Mortgage lending portfolio valuation

Government

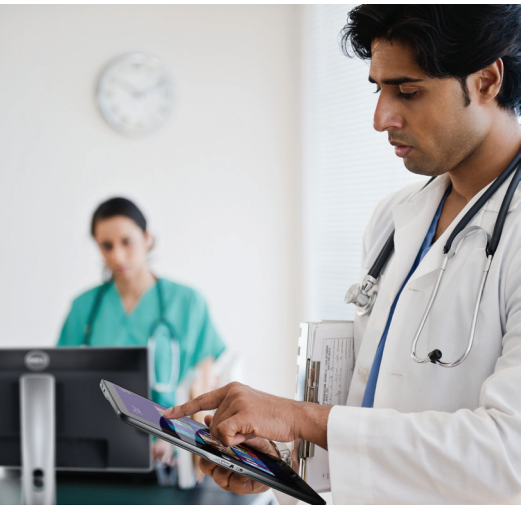
Empowering the information-driven agency

Government agencies need to leverage big data to improve productivity and derive new insights while managing risk and costs. This is a challenging proposition for many public sector entities, because budgets are limited and conventional data management architectures can't meet the technical requirements necessary to analyze diverse and petabyte-scale datasets.

Hadoop is an ideal solution here. It can provide a secure and cost-efficient enterprise data hub that allows public sector entities to store and analyze petabytes of data in various formats and from various sources, while enabling the confidence that comes with centralized oversight and security.

Hadoop use cases in government

- Data and application consolidation
- Security intelligence and fraud detection
- Data fusion and analytics for real-time and archival data



Healthcare

Improving quality and affordability
In conventional IT environments, clinical, operational and financial data are managed in data silos. Meanwhile, with the movement from paper-based to electronic health records, and with the increase in usage of machines and medical devices that produce steady streams of data, the volume of data that healthcare institutions capture and analyze has skyrocketed, while the variety of that data has grown.

The Hadoop platform allows healthcare organizations to process and manage an ever-larger influx of data in a secure and cost-effective manner to improve quality and affordability. They can leverage the platform to bring together large volumes of detailed data from diverse sources, in a variety of formats, and consolidate it into a single flexible and scalable system for long-term storage and analysis.

Hadoop use cases in healthcare
- Quality of care optimization
- Clinical quality and cost analysis
- Genome processing and DNA Population health management
- Detection of fraud and suspicious transactions

Manufacturing

Generating product and process insights
Manufacturers collect an enormous amount of data pertaining to the production of product components, the post-production performance of products, and manufacturing and supply chain processes. Today, in the era of the Internet of Things, manufacturers' data management challenges are growing in scope as products continually generate data related to their performance, functionality and quality. The challenge is not only to capture all this data, but to manage and analyze it to generate product and process insights. Hadoop provides an ideal solution to these challenges.

Hadoop use cases in manufacturing
- Proactive quality assurance
- Analysis of demand for new products and services
- Product research guided by machine-generated data
- Detection of supply chain issues
- Identification of cross-sell and upsell opportunities
- Identification of opportunities to develop new value-added services

Oil-and-gas exploration and energy utilities

Leveraging research and operational data
Oil and gas companies collect voluminous amounts of information regarding properties, geology, climate, wells and known reserves. Collecting and aggregating this data for many years of retention is a steadfast requirement. This data is a goldmine in its own right—to be treasured, enriched and exploited.

With its parallel processing capabilities and extreme scalability, the Hadoop platform can help oil and gas exploration companies make more efficient use of massive volumes of data. This is one of the keys to finding and extracting energy resources in a cost-effective manner.

Hadoop use cases in oil and gas
- Horizontal drilling enablement and optimization
- Seismic data processing
- Predicting where best to drill next
- Determining which leases to sell
- Determining which sections to acquire

Energy utilities now capture and store large amounts of data from advanced metering infrastructure, smart appliances, interactive user applications and sensors. They also make use of historical home energy data, weather data and social media data, along with disparate other types of information.

The data collection part is relatively easy. The real challenge is to consolidate and analyze this diverse range of data to answer granular questions about energy usage, predict fluctuations in demand, identify conservation opportunities and achieve various other goals of a progressive utility. Hadoop is ideally suited for these challenges.

Hadoop use cases in energy utilities

- Generate energy reports for individual consumers
- Compare energy usages among different users
- Predict sudden and temporary shifts in demand
- Gain an actionable 360-degree customer view

Retail

Cashing in on big data

To compete in the age of the Internet storefront, large retailers need scalable data management systems that integrate online and offline data so they can better understand their customers and improve the efficiency of their operations. In particular, retailers now need to connect and process data in many formats from disparate systems and sources, including the social media sites that consumers interact with.

Hadoop use cases in retail

- Enablement of a 360-degree customer view
- Generation of personalized offers
- Enablement of first in-basket analysis
- Merchandising and supply chain analysis
- Isolation of products and mixes indicative of larger baskets
- Event correlation to store traffic
- Single customer identity across all operational systems
- Loyalty program management
- Customer churn analysis

Telecommunications

Gaining business-driven insights

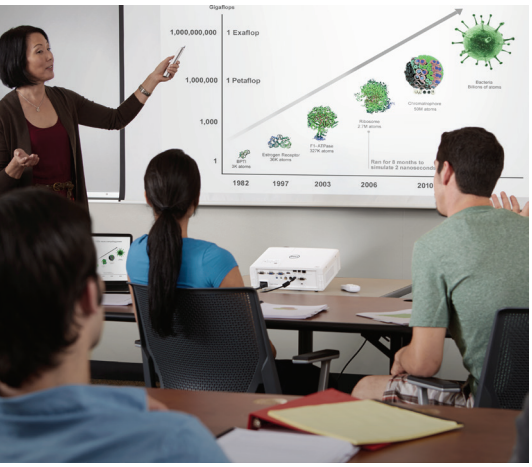
Industry challenges

Communications service providers are some of the biggest collectors of data today. Analysis of that data is one of the keys to identifying and understanding network capacity trends, reducing infrastructure costs, increasing average-revenue-per-user and preventing churn.

In traditional environments, customer information is captured in different systems. This fragmentation makes it difficult for retailers to analyze data in a holistic manner to gain a consolidated view of the customer. Hadoop allows service providers to combine information from different systems quickly and efficiently to enable large-scale data processing and analysis.

Hadoop use cases in telecommunications

- Enablement of an actionable 360-degree customer view
- Identification of affinity strengths between services and products
- Collaborative planning and forecasting
- Predictive capacity planning
- Network capacity trending and management
- Customer churn prevention
- Identification of subscriber quality of service issues
- Research and development guided by machine-generated data
- Bandwidth-hog identification



Universities and research institutions

Marrying big data and high-performance computing

The Hadoop big data platform and high performance computing (HPC) have changed the game for large-scale data analysis. Yet the two technologies are built for different purposes. HPC grew out of a need for large-scale computational speed and high performance for scientific research. Hadoop was born from the need to process large volumes of data in the Web 2.0 space. The challenge for universities is to find complementary ways to leverage Hadoop and HPC resources to accelerate time to insight on scientific investigations.

Hadoop use cases in universities and research institutions

- Big data analytics in HPC clusters
- Analyzing huge genomic datasets
- Enabling a 360-degree view of research subjects
- Data explorations to identify unpredicted correlations

Web 2.0 and technology

Gaining greater value from big data

Hadoop was inspired by papers published by Google and driven by the need to store, process and analyze massive amounts of data. Today, Web 2.0 and technology companies are challenged to gain ever-greater value from the data deluge they face on a daily basis. With this goal in mind, organizations continue to come up with new ways to leverage the Hadoop platform.

Hadoop use cases in Web 2.0 and technology

- Improving search quality through analysis of search results
- Geospatial, image and video processing
- Cyber security and fraud detection
- IP-TV subscriber recommendation engine
- House-holding and matching data across social networking and computing applications

Broad benefits

Regardless of the use cases, the Hadoop data storage and processing system offers compelling benefits for organizations that want to extract value out of huge amounts of structured, unstructured and semi-structured data. With Hadoop, you can use and store any kind of data, from any source, in its native format, and perform a wide variety of analyses and transformations on that data.

In more specific terms, Hadoop enables your organization to:

- Use and store any data in its native format without forcing transformation
- Control costs with open source software that runs on commodity hardware
- Work with industry leaders to enable a fully supported solution
- Control the rising costs and challenges of data management
- Leverage a global user and developer community that spans industries
- Scale up quickly to meet your evolving data storage and processing needs

Hadoop was originally developed by the world's largest Internet companies to capture and analyze the massive amounts of data they generate. Today, your organization can leverage the experience of these digital leaders by deploying the same platform in your environment. Even better, Hadoop allows you to start small and scale your solution to terabytes of data, or even petabytes, inexpensively.

Make your Hadoop journey with Dell and Intel

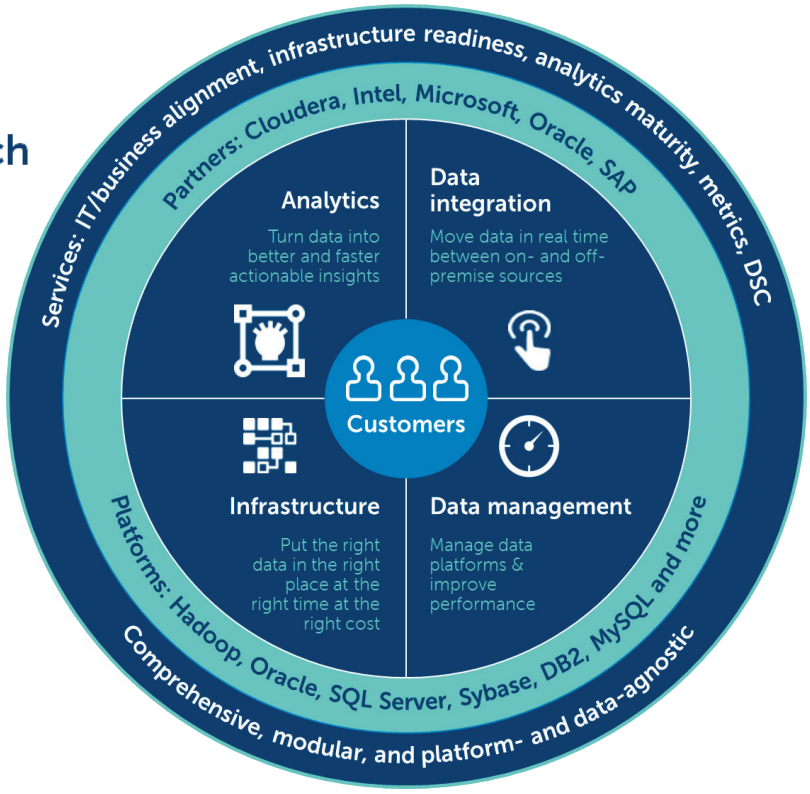
As you begin your Hadoop journey, you can look to Dell for big data expertise and the benefits of the growing portfolio of Dell™ | Cloudera® Apache™ Hadoop® big data solutions accelerated by Intel. These solutions provide end-to-end scalable infrastructure, leveraging open source technologies. They allow you to simultaneously store and process large datasets in a distributed environment for data mining and analysis on all types of data—structured, semi-structured and unstructured—and to do it all in an affordable manner.

When you partner with Dell and Intel for your Hadoop exploration and deployment, you have the confidence that comes with organizations that have worked with Hadoop for years and maintain a close working relationship with Cloudera, the leading provider of Hadoop-based software and services.

You can also look to Dell for the rest of the pieces of a complete big data solution, including unique software products for data analytics, data integration and data management. Dell offers all the tools you need to:

- **Seamlessly join structured and unstructured data.** Dell Statistica Big Data Analytics delivers integrated information modeling and visualization in a big data search and analytics platform. It seamlessly combines large-scale structured data with a variety of unstructured data, such as text, imagery and biometrics.
- **Simplify Oracle-to-Hadoop data integration.** Dell SharePlex Connector for Hadoop enables you to load and continuously replicate changes from an Oracle database to a Hadoop cluster. This toolset maintains near-real-time copies of source tables without impacting system performance or Oracle online transaction processing applications.

Dell approach to big data



- **Synchronize data between critical applications.** Dell Boomi enables you to synchronize data between mission-critical applications—on-premises and in the cloud—without the costs of procuring appliances, maintaining software or generating custom codes.
- **Easily access and merge data types.** Dell Toad Data Point can join data from relational and non-relational data sources, enabling you to easily share and view queries, files, objects and data sets.

In your explorations of your Hadoop opportunities, you can leverage the resources of a Dell Solution Center. Located in key sites around the globe, these technical centers give you the opportunity to experience Dell solutions and technology in a dedicated, hands-on environment equipped with state-of-the-art labs and teams of solution experts.

Let's get started

Ultimately, Dell, together with Cloudera and Intel, has what it takes to help you gain hands-on experience with Hadoop through a proof of concept and then take your solution into a full production environment—guided by proven reference architectures, enabled by package solutions and supported by the Dell Professional Services organization.

To learn more, visit Dell.com/Hadoop or Dell.com/BigData