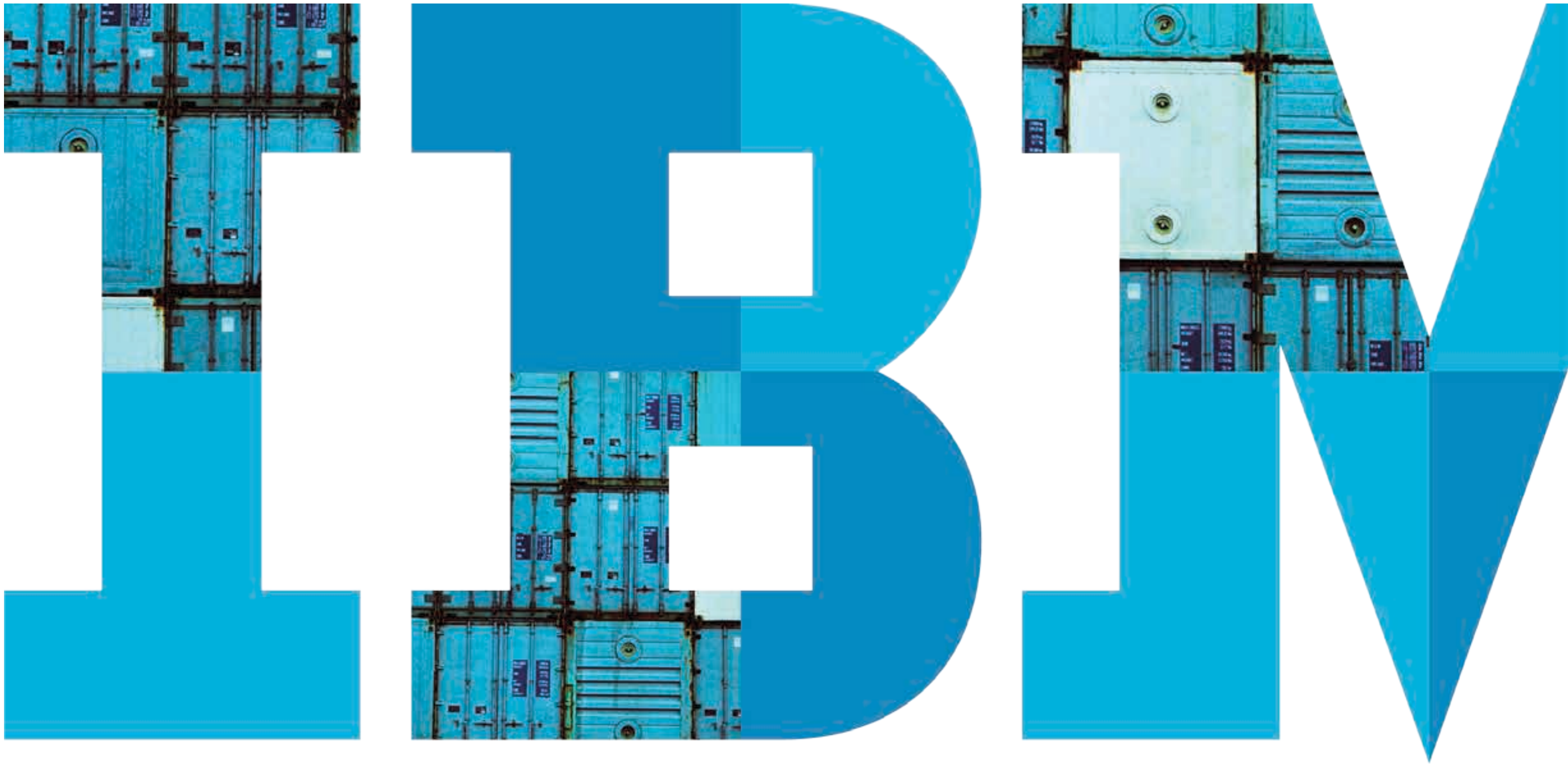
IBM Software

Hadoop in the cloud

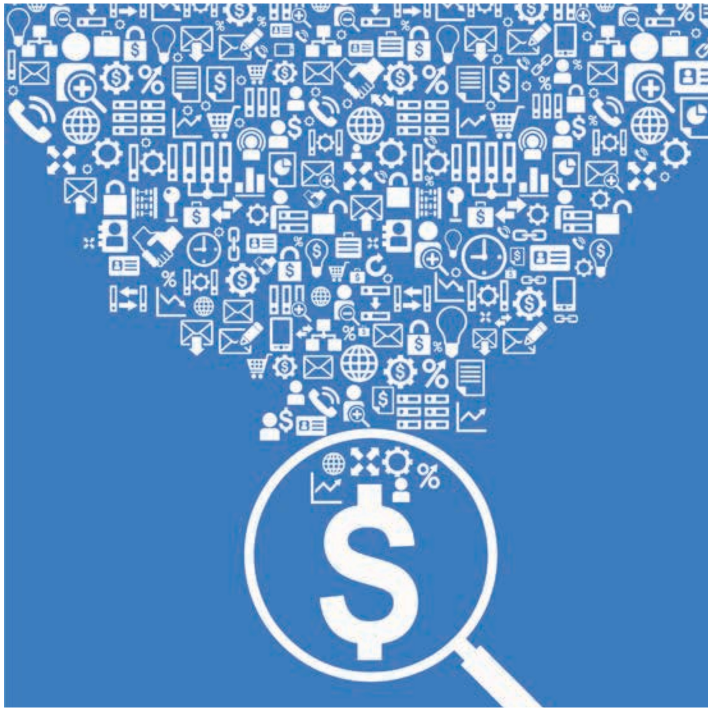
Leverage big data analytics easily and cost-effectively with IBM InfoSphere BigInsights



Introduction

One of the hottest technologies in the big data space is Apache Hadoop, an open source software framework used to reliably manage large volumes of data. Designed to scale from a single server to thousands of machines with a high degree of fault tolerance, Hadoop enables organizations to extract valuable insight from large volumes of structured, unstructured and semi-structured data.

The need for large upfront investments and concerns about flexibility, coupled with special challenges involved in evaluating the technology and developing Hadoop skills, often prevent organizations from adopting and deploying Hadoop across the enterprise. It also becomes impractical to use Hadoop on an occasional basis for high-impact projects that do not have a need for continuous processing.



There is good news, though—you can overcome these capital requirements barriers through cloud computing. The cloud model of paying only for the resources that you need—and only when you need them—supports experimentation and evaluation and is ideal for building up skills. It is also a great solution for short-term or occasional-use projects where an investment in a dedicated cluster is cost prohibitive.

What's so cool about cloud?

Cloud computing is gaining momentum for good reason. Some consider it for overall IT cost savings. Others are drawn to the promise of reduced capital expense. Still others are looking to solve pressing issues such as a chronic shortage of data center space or overly long cycles for provisioning resources (see Figure 1).

Cloud computing is the delivery of on-demand computing resources—everything from applications to data centers—over the Internet as a service. At its root, this "as-a-service" concept is simple: users can focus on their business needs and not have to worry about maintaining, improving and caring for a complex IT system. Mutually beneficial for business and IT, the cloud delivers:

- Elastic resources for quickly scaling up or down to meet spikes and lulls in demand
- Multiple payment options, from pay-as-you-go to hourly or monthly licenses
- Self-service access to technology resources

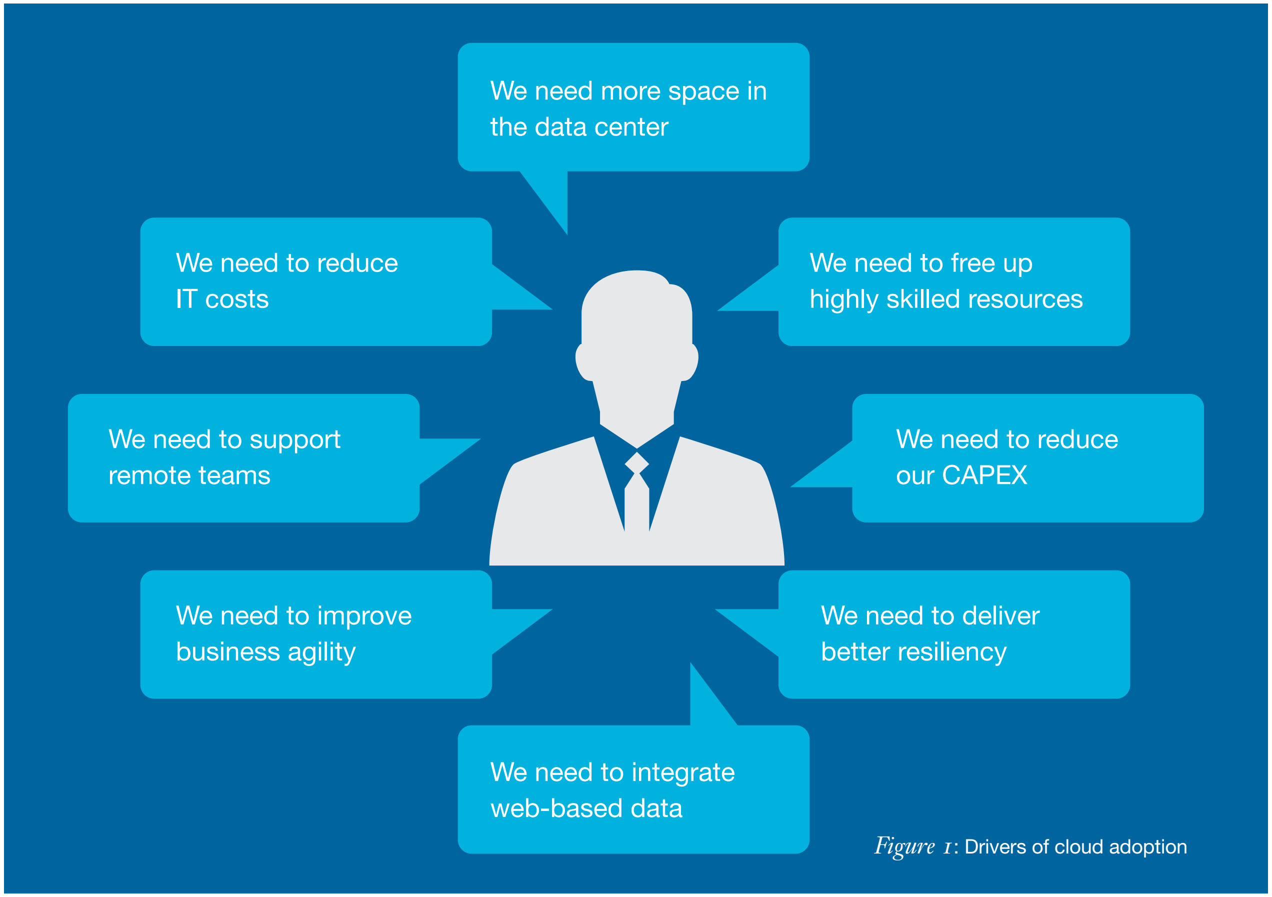

We need more space in the data center
We need to reduce IT costs
We need to free up highly skilled resources
We need to support remote teams
We need to reduce our CAPEX
We need to improve business agility
We need to deliver better resiliency
We need to integrate web-based data

Figure 1: Drivers of cloud adoption

This e-book explores how you can use IBM-enhanced Hadoop capabilities in the cloud to cost-effectively deploy deep analytics for all users in your organization—opening up the benefits of big data to everyone.

Cloud and analytics: The new growth engine for business

In the view of IBM, cloud is more than just a way to manage costs or get services faster—it is a critical path to business growth. Cloud and big data analytics offer the potential to place information, insights and decision making at people's fingertips, at the right time and place.

Because cloud computing offers access to unlimited computing power and easier ways to process large amounts of data, it's an ideal deployment model for big data analytics. IBM has combined cloud computing and big data analytics in a wide range of industries, resulting in benefits such as:

- The discovery of life-changing medicines
- More accurate prediction of weather patterns
- Innovative energy-saving techniques
- Insight into security anomalies
- More efficient ways to use and conserve water
- Deeper insight into customer preferences and trends
- Real-time feedback on marketing campaigns

The shift toward cloud computing is also a response to the realization that big data and analytics must take a more central role in today's business world, becoming an engine that helps drive the business forward. **Organizations need to transition from passive, siloed "systems of record" designed around discrete pieces of information to "systems of engagement,"** which are more decentralized, incorporate technologies that encourage peer interactions and often leverage cloud technologies to enable those interactions.

The differences between the two systems—and the ramifications of those differences—are significant. For example, in order to integrate systems and support enhanced collaboration (a central tenet of systems of engagement), a company needs to deploy appropriate platform technologies. Formats include Software as a Service (SaaS), Platform as a Service (PaaS) or Infrastructure as a Service (IaaS) offerings, and can be deployed on the public cloud, a private cloud or a hybrid model.

Enhancing Hadoop in the cloud with IBM InfoSphere BigInsights

Cloud computing enables you to overcome the capital requirements of Hadoop. But open source Hadoop lacks enterprise-grade management technology and performance and may require organizations to learn or acquire new skills. Organizations can overcome these barriers by using IBM® InfoSphere® BigInsights™, an enterprise-ready distribution of Hadoop. In addition to the Hadoop technology, InfoSphere BigInsights delivers unique value that is designed to address the challenges of modern enterprise IT:

- **Extended Hadoop:** InfoSphere BigInsights is based on 100 percent open source Hadoop. It extends Hadoop with enterprise-grade technology including administration and integration capabilities, visualization and discovery tools as well as security, audit history and performance management.
- **Increased performance:** An average 4 times performance gain over open source Hadoop.[1]
- **Usability:** InfoSphere BigInsights is optimized for a wide range of roles, including integration developers, administrators, data scientists, analysts and line-of-business contacts.
- **Integrated with IBM Watson™ Foundations big data platform:** InfoSphere BigInsights comes bundled with search and streaming analytics capabilities.
- **Analytics:** Built-in Hadoop analytics capabilities for machine data, social data, text and Big R enable you to locate actionable insights from data in the Hadoop cluster rather than having to move the data around.

This combination—open source Hadoop and the value-add enterprise features from IBM—can be deployed on the cloud. IBM provides pre-built images and/or templates for rapid deployment of Hadoop clusters in the cloud, and multi-cloud templates for InfoSphere BigInsights are also available through RightScale. Deploying an InfoSphere BigInsights cluster on the cloud also frees users from sourcing (and paying for) extra equipment and racks.

The many roles of analytics in the cloud

The IaaS and PaaS options will be used by developers, but business leaders will see a significant impact too. For example, clients deploying InfoSphere BigInsights on the cloud have reduced hardware and software costs, avoided future expansion costs, simplified development and management processes, and realized dramatic performance improvement. Check out the table for more examples of how cloud analytics are delivering real-world benefits.

Industry	Cloud analytics use case	Benefits
Retail	Support growing data volumes and analyze customer data in more efficient ways to acquire deeper, more valuable insights.	Reduced costs and improved marketing effectiveness.
Healthcare	Analyze millions of patient records and perform statistically guided decision support to lower diagnostic errors and improve the quality of care.	Improved outcomes and better insights into treatment trends and relationships
Energy and utilities	Implement proactive resource optimization and allocation, perform asset management and maintenance optimization.	More efficient resource utilization and potentially less downtime or capacity shortfalls
Banking	Identify customer patterns from log data to improve customer insight and provide better-targeted offers and services. Create a one-stop shop for data discovery.	Increased up-sell and cross-sell opportunities; more granular customer demographic segmentation

IBM Watson Foundations:
Complete cloud analytics capabilities

InfoSphere Big Insights draws additional strength from its lineage as part of the IBM Watson Foundations big data portfolio. IBM Watson Foundations delivers a full range of capabilities to help you meet your big data and analytics goals:

- **Real-time analytics:** Dynamically update business rules and processes based on what's happening right now. Analyze data in motion for real-time insights.
- **Real-time application development:** Quickly ingest, analyze and correlate data as it arrives from thousands of real-time sources. Easily build applications with drag operators, visual editors and performance monitoring. Dynamically add new data sources. Create, edit, visualize, test, debug and run applications in the cloud.
- **Analytic toolkits and accelerators:** Deploy deep analytics developed by IBM Research, such as geospatial, time series, R analysis, text analytics and much more.

Performance, scalability and deep analytics make IBM big data solutions exceptional. They deliver:

- **More volume:** Handles up to 10 times more records per second on the same hardware compared to other open source and complex event processing (CEP) vendors[2]
- **More data variety:** Analytics and powerful modeling on any and all data types
- **More velocity:** One-tenth to one-thousandth the latency compared to other open source and CEP vendors[3]

IBM Watson Foundations also contributes key big data and analytics capabilities optimized for the cloud (see Figure 2).

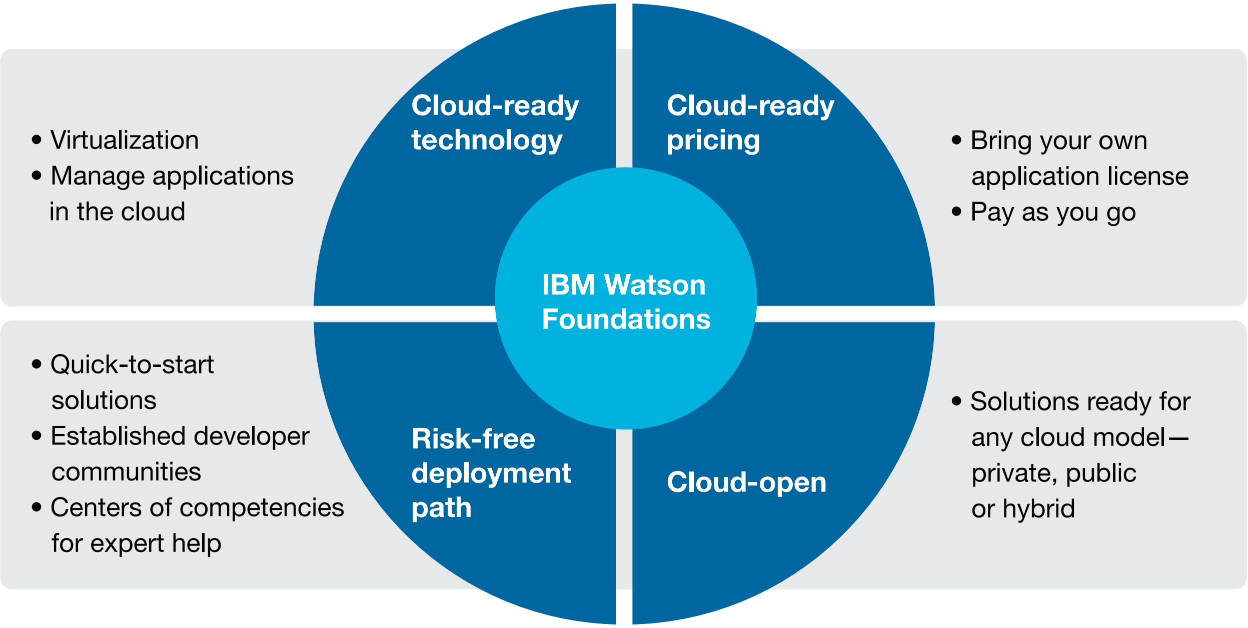

• Virtualization
• Manage applications in the cloud
Cloud-ready technology
Cloud-ready pricing
• Bring your own application license
• Pay as you go
IBM Watson Foundations
• Quick-to-start solutions
• Established developer communities
• Centers of competencies for expert help
Risk-free deployment path
Cloud-open
• Solutions ready for any cloud model—private, public or hybrid

Figure 2. Cloud capabilities provided by the IBM Watson Foundations big data platform.

A portfolio of solutions that work together

As part of IBM Watson Foundations, InfoSphere BigInsights works together with other IBM big data solutions such as InfoSphere Streams to deliver exceptional results.

For example, Wimbledon Championships served up an ace performance using InfoSphere Streams and InfoSphere BigInsights, breaking new ground with 433 million page views serving up a total of 155 TB of data—equivalent to over 35 years' worth of CD-quality audio recording. More detailed and comprehensive analysis of current and past data enabled the tournament organizer to create more interesting and engaging content for the 19.7 million unique users.

The IBM system gathered large volumes of data in real time from on-court sensors and scorers plus social media from off-court analysts and fans around the globe, and then integrated it with other sources of structured and unstructured data for distribution to analytical tools, websites, mobile apps and broadcasters. The real-time analysis of match data revealed winning patterns; analytics were also used to predict demand, enabling the cloud infrastructure to automatically adjust resources.

InfoSphere BigInsights for the cloud allows organizations to respond faster to changing business environments by analyzing larger volumes of data more cost-effectively. It enables organizations to analyze all data in its native format to add real-world information to decision processes. Organizations can use it to scale to petabytes of data and thousands of users with near-linear processor scalability—all on a reliable and secure platform.

Resources

In this era of big data, you need solutions that allow you to easily and cost-effectively unlock the value of enterprise data. Many analytics solutions leave users frustrated or disappointed because they can't act handle today's big data volumes, take too long to deploy or require huge new upfront investments.

It's time for a new approach. Analytics in the cloud with InfoSphere BigInsights allows you to easily and economically tap into the power of Hadoop and big data. To learn more about how you can take advantage of InfoSphere BigInsights and IBM cloud offerings, visit these resources:

IBM InfoSphere BigInsights overview

InfoSphere BigInsights Quick Start Edition

IBM Watson Foundations

IBM Cloud Computing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
April 2014

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] 4 times is approximate value. Testing involved the SWIM benchmark (https://github.com/SWIMProjectUCB/SWIM) and jobs derived from production workload traces. Testing was conducted in controlled laboratory conditions. See "STAC Report: Comparison of IBM InfoSphere BigInsights Enterprise Edition with Apache Hadoop using SWIM." www.stacresearch.com/node/15370

[2, 3] IBM InfoSphere Streams v3.0 Performance Report. February 2013. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500012717&S_CMP=is_dwwp14_ppo