# IBM BigInsights for Apache Hadoop

*Efficiently manage and mine big data for valuable insights*

Highlights:

- Enterprise-ready Apache Hadoop–based platform for data processing, warehousing and analytics
- Advanced analytics for structured, semi-structured and unstructured data
- Professional-grade visualization, development and administration tooling to boost productivity
- Application accelerators that help speed implementation and accelerate time-to-value
- Integration with proven IBM offerings as well as third-party solutions

## Tame big data

IBM® BigInsights for Apache™ Hadoop® enables organizations to turn large, complex data volumes into insights by addressing a multitude of business challenges. At a high level, these challenges can be broken down into three main categories: operational efficiency, advanced analytics, and exploration and discovery.

### Operational efficiency

To more effectively handle the performance and economic impact of growing data volumes, architectures incorporating different operational characters can be used together. For example, large amounts of cold data in the data warehouse can be archived to an analytics environment rather than to a passive store.

BigInsights helps improve operational efficiency by modernizing—not replacing—the data warehouse environment. It can be used as a query-able archive, enabling organizations to store and analyze large volumes of poly-structured data without straining the data warehouse. As a preprocessing hub—also referred to as a "landing zone" for data—BigInsights helps organizations explore their data, determine the high-value assets and extract that data cost-effectively. It also supports ad hoc analysis of large amounts of data for exploration, discovery and analysis.

### Advanced analytics

In addition to increasing operational efficiency, some organizations are looking to perform new, advanced analytics but lack the proper tools. With BigInsights, analytics is not a separate step performed after data is stored; instead, BigInsights, in combination with InfoSphere Streams, enables real-time analytics that can leverage historic models derived from data being analyzed at rest. BigInsights includes advanced text-analytic capabilities and prepackaged accelerators. Organizations can use these pre-built analytic capabilities to understand the context of text in unstructured documents, perform sentiment analysis on social data or derive insight from a wide variety of data sources.

### Exploration and discovery

The explosive growth of big data may overwhelm organizations, making it difficult to uncover nuggets of high-value information. BigInsights helps build an environment well suited to exploring and discovering data relationships and correlations that can lead to new insights and improved business results. Data scientists can analyze raw data from big data sources alongside data from the enterprise warehouse and several other sources in a sandbox-like environment. Subsequently, they can combine any newly discovered high-value information with other data to help improve operational and strategic insights and decision making.

The bottom line: with BigInsights, enterprises can finally get their arms around massive amounts of untapped data and mine it for valuable insights in an efficient, optimized and scalable way.

## Bring Hadoop to the enterprise

BigInsights for Hadoop combines open-source Apache Hadoop with IBM innovations to deliver massive scale-out data processing and analysis with built-in resiliency and fault tolerance. IBM has built simplified administration and management capabilities, rich developer tools and powerful analytic functions—reducing the complexity of getting started with Hadoop.

One of the biggest challenges in building applications using open-source or third-party Hadoop distributions is the high level of skill involved. BigInsights solves the problem by making it easy for the two largest populations of data processing skills available—spreadsheet users and SQL programmers—to create applications and get insights.

### Big SQL

Big SQL uses a massively parallel processing (MPP) SQL engine directly on the physical Hadoop Distributed File System (HDFS) cluster rather than using Map-Reduce, vastly improving performance and SQL execution capabilities over Apache Hive 12. Big SQL leverages standard SQL to allow users to access big data in the same way they leverage other relational data. BigInsights also provides a built-in interactive dashboard for end-user interaction with big data out of the box and it integrates via Big SQL seamlessly into IBM Cognos® Business intelligence for interactive dashboards and activities.

### The power of Hadoop

**BigInsights enhances open-source Hadoop with the enterprise-class functionality and integration necessary to meet critical business requirements. Organizations can run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure leverages the Hadoop MapReduce framework to tackle very large data sets by breaking up the data across many nodes and coordinating data processing across a massively parallel environment. After the raw data has been stored across the distributed cluster, the systems can efficiently handle queries and data analysis.**

**Performance** Benchmark tests indicate that Big SQL executes queries **20 times faster, on average**, over Apache Hive 12 with performance improvements ranging up to 70 times faster for individual queries.

**Comprehensive SQL support** Big SQL 3.0 has successfully run **ALL 99 TPC-DS queries** and **ALL 22 TPC-H queries without modification**. To contrast, Apache Hive 12 executes only 43 of the 99 TPC-DS queries without modification.

**Row and column access** Big SQL enables row and column access control, or "fine-grained control" consistent with functionality found in an RDBMS.

**Federated data access** Big SQL can access data from more than BigInsights. Its federated access allows users to send distributed requests to multiple data sources within a single SQL statement.

Administrators start with a GUI-driven installation tool that guides them to specify which optional components to install and how to configure the platform. Installation progress is reported in real time, and a built-in health check is designed to automatically verify the success of the installation. These advanced installation features minimize the amount of time needed for installation and tuning, freeing administrators to work on other critical projects.

Once the Hadoop cluster is in place, robust job management features give organizations control of BigInsights jobs, user roles, security and key performance indicator (KPI) monitoring. Technical staff can easily direct job creation, submission and cancellation; they can also stay informed of workload progress through integrated job status dashboards, logs and monitors that provide details on configuration, tasks, attempts and other critical information. In addition, BigInsights provides administration features for Hadoop Distributed File System (HDFS), IBM GPFS™ File Placement Optimizer (FPO), big data applications and MapReduce jobs, and cluster management.

As shown in Figure 1, BigInsights for Hadoop provides several enterprise capabilities. The following sections detail each area of these capabilities.

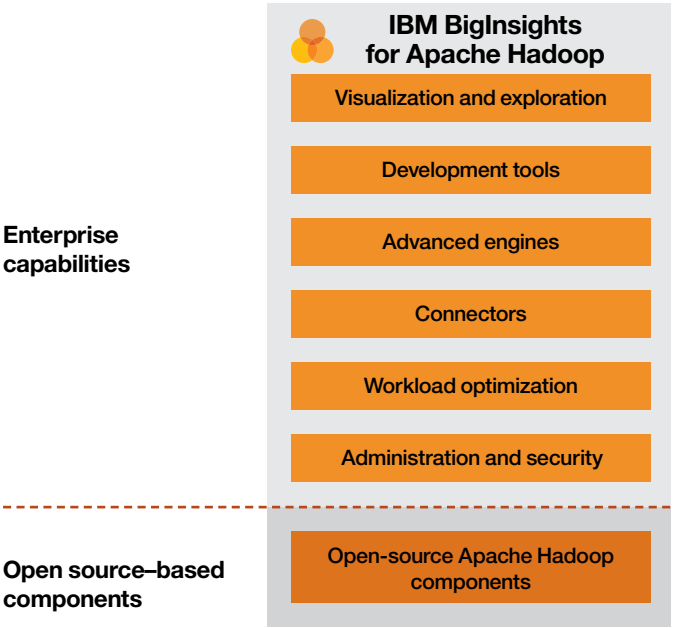


*Figure 1.* BigInsights adds enterprise capabilities to open-source components.

### Try BigInsights at no cost

**BigInsights Quick Start Edition is a no-charge, downloadable, nonproduction version of BigInsights. It gives you the chance to explore Hadoop without data capacity or time limitations. To download your Quick Start Edition today, visit: ibm.com/software/data/infosphere/biginsights/quick-start**

## Visualization and exploration

BigInsights enables exploration and ad hoc analysis of all data stored in the platform, as well as enabling users to visualize it in several ways.

### BigSheets, data exploration and dashboards

BigSheets is a browser-based, spreadsheet-style tool that enables data scientists and business users to explore, manipulate and analyze big data.

BigSheets can help business users perform the following tasks:

- Integrate and explore large amounts of data in different formats and structures.
- Extract and enrich data using text analytics.
- Explore and visualize data with charts and pivot tables.

BigInsights also comes with a centralized dashboard that allows business analysts to get insights from their data and view large-scale analytics results. Administrators can use the dashboard to monitor key performance metrics of their BigInsights for Hadoop cluster.

## Development tools

BigInsights uses a familiar, Eclipse-based development environment for building and deploying applications. It provides editors for Hadoop components such as Java™ MapReduce, Hive and Pig. It also provides a programmer interface for Big SQL, Oozie Workflows and Text Analytics.

BigInsights also comes with unified development lifecycle tooling, which enables users to sample data from Hadoop, bring it to the development environment, and develop, test and deploy applications to the cluster.

## Advanced engines and accelerators

BigInsights includes a sophisticated set of analytics tools and capabilities at no additional charge. Out of the box, organizations can quickly begin uncovering patterns in their data and build powerful, custom analytic applications that deliver results and insights tailored to specific business needs.

### Advanced text analytics

BigInsights includes a powerful text analytics engine developed by IBM Research. Using a comprehensive library of rules or by developing their own custom rules, users can quickly extract and identify items of interest in documents and messages, including people, email addresses, street addresses, phone numbers, URLs, joint ventures, alliances and more.

### Social Data Analytics Accelerator

The Social Data Analytics Accelerator enables users to analyze various types of social media data to gain key insights to support BI. It can capture vital consumer intelligence including sentiment, purchase intent and product/service ownership as well as demographic attributes such as gender, location, parental status, marital status, employment, interests, current customer of, products owned and product interest. Organizations can leverage these attributes to build applications such as lead generation, customer retention/churn reduction, customer acquisition and targeted marketing campaigns.

### Machine Data Analytics Accelerator

The Machine Data Analytics Accelerator can ingest, parse and extract a variety of machine data from sources such as log files, smart devices and telemetry, and help process that data in minutes instead of days and weeks. Organizations gain insights into operations, transactions and system behavior. The resulting information can be used to proactively boost operational efficiency, troubleshoot or identify root causes of problems and investigate incidents, which helps the company avoid service degradation or outages.

## Connectors

Big data technologies can play an important role in the enterprise information supply chain, but only if they are deeply and tightly integrated with existing systems. IBM recognizes this and developed BigInsights with high-speed connectors for data of all types (structured, unstructured and streaming) and sources (data warehouse, social media, log data and so on). The built-in integration connectors can move data to structured systems as well as to the Hadoop file system, while BigInsights can directly ingest unstructured data.

BigInsights provides connectors to IBM DB2® database software, the IBM PureData™ Systems family of data warehouse appliances, IBM Netezza appliances, IBM InfoSphere Warehouse and the IBM Smart Analytics System. These high-speed connectors help simplify and accelerate data manipulation tasks. Standard Java Database Connectivity (JDBC) connectors make it possible for organizations to quickly integrate with a wide variety of data and information systems including Oracle, Microsoft® SQL Server, MySQL and Teradata.

In addition, IBM InfoSphere DataStage® includes a connector that enables BigInsights data to be leveraged within an InfoSphere DataStage extract/transform/load (ETL) or in an extract/load/transform (ELT) job.

## Workload optimization

BigInsights provides several features that help increase performance, as well as enhance its adaptability and compatibility within an enterprise environment.

### Scheduler for adaptable workflow allocation

Not all workloads have the same priority. The BigInsights Scheduler provides an adaptable workflow allocation scheme for MapReduce jobs that optimizes processing based on a user-chosen policy. The scheduler is an extension to the Hadoop Fair Scheduler, which is designed to, over time, allot all jobs an equitable share of cluster resources.

### Adaptive MapReduce for job acceleration

Jobs running on Hadoop can end up creating multiple small tasks that consume a disproportionately large amount of system resources. To combat this, IBM invented a technique called Adaptive MapReduce that is designed to speed up small jobs by changing how MapReduce tasks are handled without altering how jobs are created. Adaptive MapReduce is transparent to MapReduce operations and Hadoop application programming interface (API) operations.

## Administration and security

Stringent enterprise security requirements must extend to big data, just as they apply to all other enterprise information resources. BigInsights delivers several sophisticated options that help ensure data security and privacy.

### Authentication

Administrators have the option to choose flat file, Lightweight Directory Access Protocol (LDAP) or Pluggable Authentication Modules (PAM) for the BigInsights web console. With LDAP authentication, the BigInsights installation program will communicate with an LDAP credentials store for authentication. Administrators can then provide access to the BigInsights console based on role membership, making it easy to set access rights for groups of users.

### Roles

BigInsights provides four levels of user roles: system administrator, data administrator, application administrator and non-administrative user. Access to data and features depends on the user's assigned role.

### Auditing and Security

MapReduce jobs can be run under designated account IDs, which helps tighten security, access control and auditing. And integration of BigInsights with IBM InfoSphere Guardium® data security software helps organizations to manage the security and auditing needs of Hadoop the same way they manage traditional structured data sources.

BigInsights also supports Kerberos service-to-service authentication protocol, increasing security strength to prevent middle man attacks.

## Enhanced enterprise integration

### IBM Watson Explorer

BigInsights includes a limited-use license for Watson Explorer, which helps organizations discover, navigate and visualize vast amounts of structured and unstructured information across enterprise systems and data repositories. It also provides a cost-effective and efficient entry point to explore the value of big data technologies through a powerful framework for developing applications that leverage existing enterprise data.

### InfoSphere Streams

BigInsights includes a limited-use license of InfoSphere Streams, which enables real-time, continuous analysis of data on the fly. InfoSphere Streams is an enterprise-class stream-processing system that can extract actionable insights from data in motion while transforming data and transferring it to BigInsights at high speeds. This enables organizations to capture and act on business data in real time—rapidly ingesting, analyzing and correlating information as it arrives—and fundamentally enhance processing performance.

### Cognos Business Intelligence

BigInsights includes a limited-use license for Cognos Business Intelligence, which enables business users to access and analyze the information they need to improve decision making, gain better insight and manage performance. Cognos Business Intelligence includes software for query, reporting, analysis and dashboards, as well as software to gather and organize information from multiple sources.

### InfoSphere Master Data Management

For users performing customer analytics, BigInsights leverages the probabilistic matching engine of InfoSphere Master Data Management to match and link customer information directly in Hadoop, at high speeds. A unique identifier for each customer ensures analytics are performed on more accurate and information.

### Conclusion

BigInsights for Hadoop is 100 percent Apache Hadoop Open Source and includes enterprise grade capabilities to support all big data use cases. IBM enhances the Hadoop experience with high availability, training, support and services required to ensure successful deployment and ROI.

## For more information

To learn more about the IBM BigInsights for Apache Hadoop, please contact your IBM sales representative or IBM Business Partner, or visit:
**ibm.com**/software/data/infosphere/biginsights



IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
March 2015



IMD14385-USEN-04