IBM InfoSphere BigInsights Enterprise Edition

Efficiently manage and mine big data for valuable insights

Highlights

- 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 popular IBM offerings as well as third-party solutions
- Enterprise-ready Apache Hadoop-based analytics platform
- Support for multiple Hadoop distributions for maximum flexibility

Tame big data

Many companies are seeing dramatic growth in the variety, velocity and volume of information being generated by their businesses and are struggling with how to manage big data: vast and diverse quantities of traditional structured data as well as semi-structured and unstructured data. Big data offers tremendous potential for deep insights that drive fast, clear and nuanced decision making, but to take advantage of this resource, organizations need data management and analysis tools that are effective at a completely different level than ever before.

IBM® InfoSphere® BigInsights™ Enterprise Edition enables organizations to turn large, complex data volumes into insight 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. Data may also be integrated through a common repository on the way to the warehouse.

InfoSphere BigInsights helps improve operational efficiency by augmenting—not replacing—the data warehouse environment. It can be used as a query-able archive, enabling organizations to store and analyze large volumes of multi-structured data without straining the data warehouse. As a preprocessing hub or a "landing zone" for data, InfoSphere BigInsights helps organizations explore their data, determine the high-value assets and extract that data cost-effectively. In addition, InfoSphere BigInsights supports ad hoc analysis of large amounts of data to help organizations explore, discover and analyze enterprise data very quickly.

Advanced analytics

In addition to increasing operational efficiency, some organizations are looking to perform new, advanced analytics but lack the proper tools. With InfoSphere BigInsights, analytics is not a separate step performed after data is stored; instead, InfoSphere BigInsights acts in combination with InfoSphere Streams to enable real-time analytics that leverage historic models derived from data analyzed at rest.

InfoSphere BigInsights includes several pre-built analytic modules and prepackaged accelerators that organizations can use to understand the context of text in unstructured documents, perform sentiment analysis on social data or derive insight out of data from a wide variety of sources.

Exploration and discovery

The diverse sources and types of big data, combined with the explosive growth in data volume, may overwhelm organizations, making it difficult to uncover nuggets of high-value information. InfoSphere 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 sample data from the enterprise warehouse in a sandbox-like environment. Subsequently, they can move any newly discovered high-value data into the enterprise data warehouse and combine it with other trusted data to help improve operational and strategic insights and decision making.

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

Bring Hadoop to the enterprise

InfoSphere BigInsights combines Apache Hadoop with IBM innovations to deliver massive scale-out data processing and analysis with built-in resiliency and fault tolerance. IBM has built functionalities on top of Hadoop, adding 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; with InfoSphere BigInsights, developers and other users can easily build applications and get insights with their existing knowledge.

Administrators start with a GUI-driven installation tool and guided installation that allows 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 solution is in place, robust job management features give organizations fine-grained control of InfoSphere BigInsights jobs. Technical staff can easily direct job creation, submission and cancellation; they can also stay informed of workload progress through integrated job status displays, logs and counters that provide details on configuration, tasks, attempts and other critical information. In addition, InfoSphere BigInsights provides administration features, including Hadoop Distributed File System (HDFS) and MapReduce administration, cluster and server management, role-specific views and the ability to view HDFS file content.

Building on the power of Hadoop

InfoSphere BigInsights takes open-source Hadoop and adds the 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. Once the raw data has been stored across the distributed cluster, queries and analysis of the data can be handled efficiently, with dynamic interpretation of the data format at read time.

Boost flexibility, consumability and manageability

Version 2.0 of InfoSphere BigInsights includes enhanced visualization, monitoring and development tools that enable various roles across an organization to collaboratively unlock the value of data. It brings relevant data together for analysis and avoids organizational silos.

Business analysts

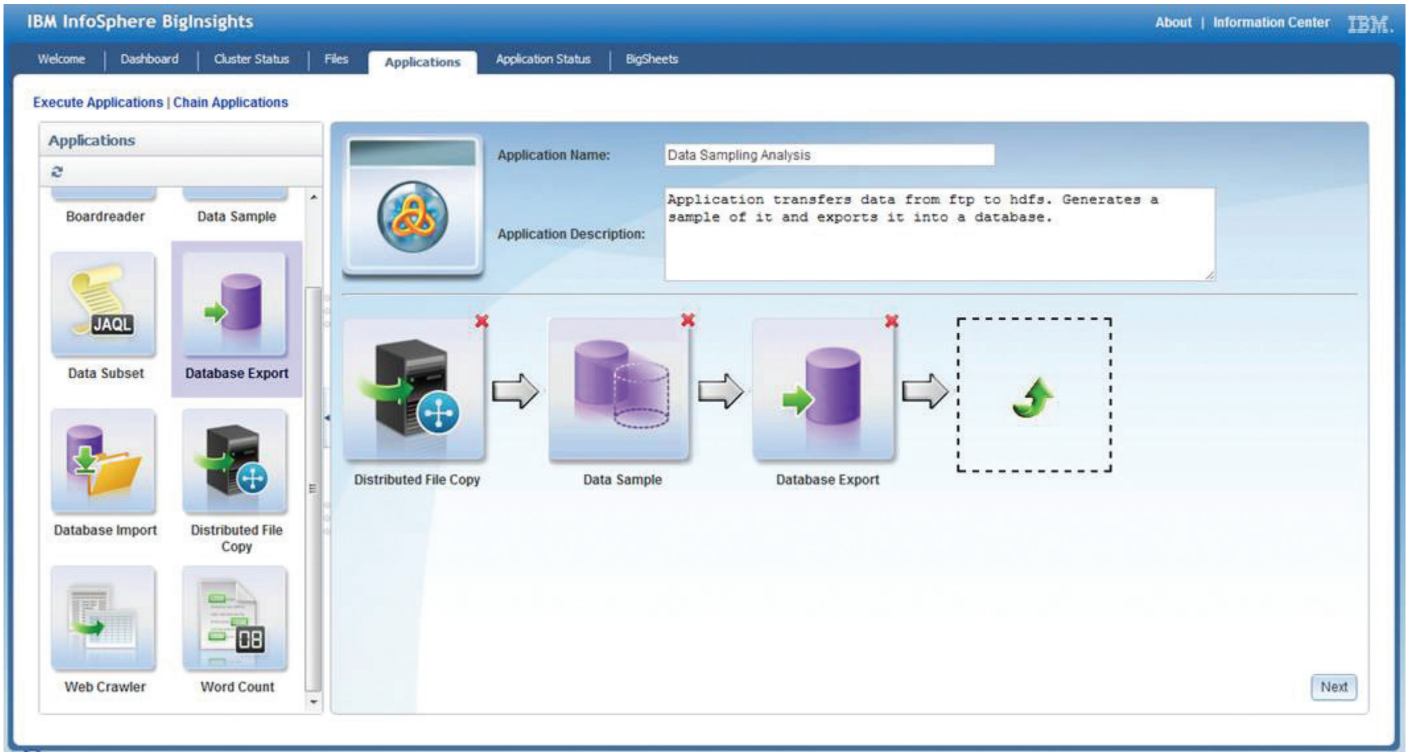
InfoSphere BigInsights V2.0 introduces a centralized dashboard that leverages a new charting engine. Business analysts can use the dashboard to get insights into their data, view analytic application results and monitor metrics. Application linking enables business analysts to compose new applications from existing ones and from IBM BigSheets, a spreadsheet-like visualization tool, without learning special programming skills. Business analysts can also use application linking to invoke analytics applications from the web console (see Figure 1).

Usability enhancements enable business analysts to view BigSheets data flows between and across data sets to quickly navigate and relate analysis and charts.

Data scientists

Text analytics enhancements add support for modular extractors. By enabling the reuse of common extractors, dictionaries, rules and tables, these capabilities help improve the usability and development productivity of text analytics applications. Data scientists can also manage the text analytics lifecycle with new capabilities that allow for sampling and subsetting data; developing, testing and publishing text analytics applications from the InfoSphere BigInsights tools for Eclipse; and deploying, executing and monitoring text analytics applications from the web console.

In addition, integration with the R open-source programming language makes it possible for data scientists to execute an ad hoc R job directly from the InfoSphere BigInsights web console.



Application Name: Data Sampling Analysis

Application Description: Application transfers data from ftp to hdfs. Generates a sample of it and exports it into a database.

Figure 1: Application linking for business analysts.

Administrators

Version 2.0 of InfoSphere BigInsights supports agile administration through new monitoring capabilities. The centralized dashboard view lets administrators visualize key performance indicators including CPU, disk, memory and network usage for the cluster, as well as for Apache data services such as HDFS, HBase, Zookeeper and Flume, and Apache application services including MapReduce, Hive and Oozie (see Figure 2). Administrators also get enhanced status information and control over applications and major cluster capabilities.

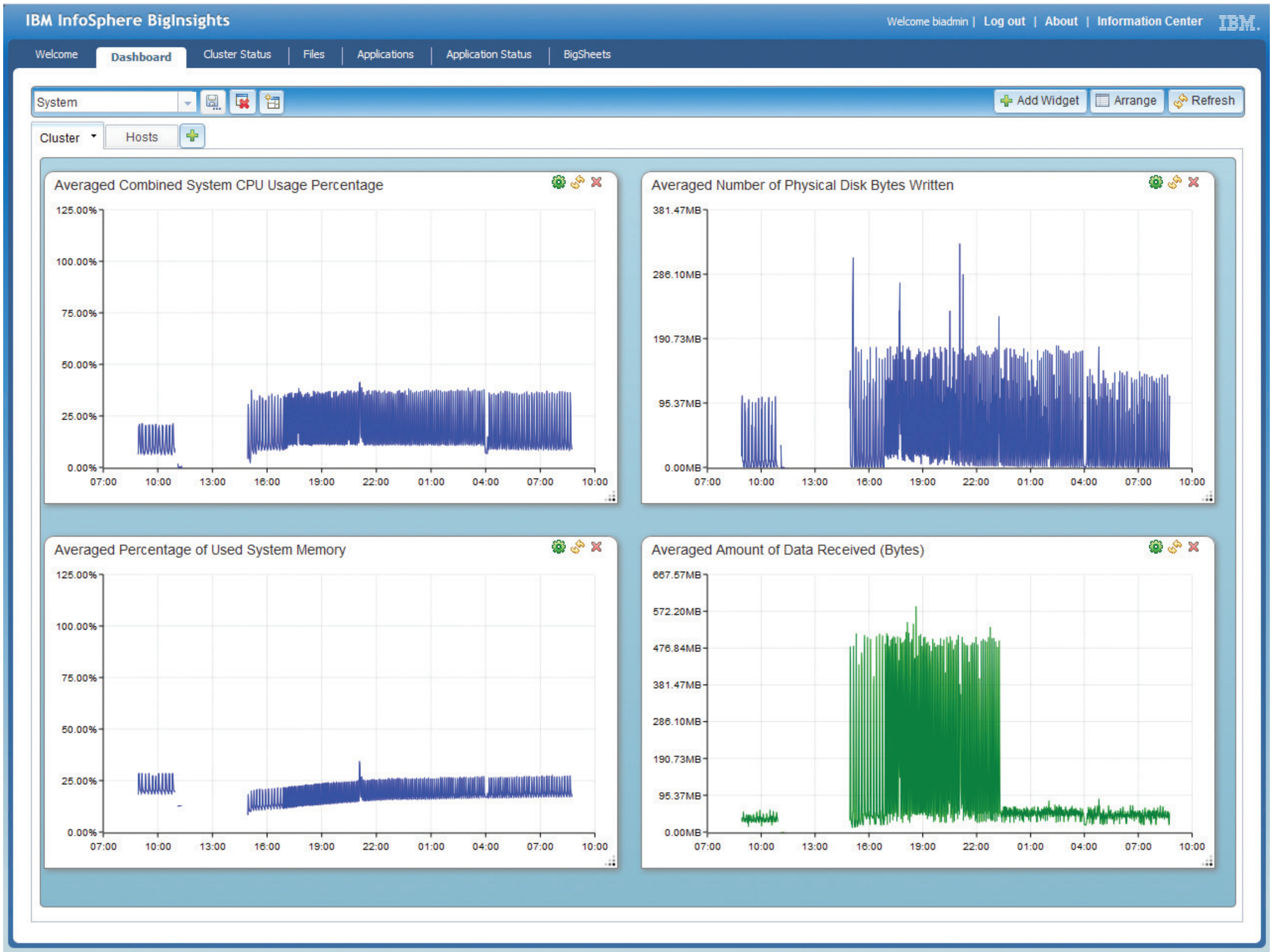


Figure 2: Enhanced monitoring capabilities for administrators.

Developers

Developers have access to professional-grade application development tooling that makes it simple to build applications without being a Hadoop expert. The unified tooling enables developers to sample data and define, test and deploy analytics applications from the Eclipse tools, as well as to administer, execute and monitor deployed applications from the web console. This tooling includes the following features:

- New and enhanced editors, such as a workflow editor that greatly simplifies the creation of complex Oozie workflows with a consumable interface; a Pig editor with content assist and syntax highlighting; and an improved Jaql editor with extended support for Jaql syntax, extended content assist, and improved execution feedback
- Support for new application types with enablement of BigSheets macro and reader development
- Advanced text analytics development, including support for modular rule sets
- Enhanced scope of development artifacts during the deployment phase, including artifacts for text analytics; scripts for Jaql, Hive SQL and Pig; and BigSheets macros and readers

Dive deep into big data with analytic and application accelerators

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

Advanced text analytics

InfoSphere BigInsights includes the powerful text analytics engine developed by IBM. Using a comprehensive library of rules or their own custom rules, developers can quickly query 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

New to InfoSphere BigInsights V2.0, the social data analytics accelerator can be used to efficiently analyze customer information from various types of social media data. This analysis supports enhanced insights for effective, targeted marketing campaigns and timely product marketing decisions—helping organizations gain competitive intelligence and build customer retention and new customer acquisition programs.

Machine Data Analytics Accelerator

Also new in InfoSphere BigInsights V2.0, 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. By using the machine data analytics accelerator, organizations gain insights into operations, customer experience, transactions and behavior. The resulting information can be used to proactively boost operational efficiency; troubleshoot problems and investigate security incidents; and monitor end-to-end infrastructure to avoid service degradation or outages.

Add security to big data analysis

Stringent enterprise security requirements must extend to big data, just as they apply to all other enterprise information resources. InfoSphere 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 no authentication for the InfoSphere BigInsights web console. With LDAP authentication, the InfoSphere BigInsights installation program will communicate with an LDAP credentials store for authentication. Administrators can then provide access to the InfoSphere BigInsights console based on role membership, making it easy to set access rights for groups of users.

Roles

InfoSphere 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

MapReduce jobs can be run under designated account IDs, which helps tighten security, access control and auditing. In addition, a new key store enhances password security. And integration of InfoSphere BigInsights with IBM InfoSphere Guardium® data security software helps organizations manage the security of Hadoop-based and traditional structured data.

Maximize performance, streamline job handling

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

InfoSphere BigInsights Scheduler for adaptable workflow allocation

Not all workloads have the same priority. The InfoSphere 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 guarantee that, over time, all jobs get an equitable share of cluster resources.

BigIndex for large-scale indexing

BigIndex helps make Hadoop-based indexing easy by including it as a native capability in InfoSphere BigInsights. BigIndex delivers low-latency, full-text search capabilities for big data. Indexes can be built, scanned and queried using the BigIndex module as part of a workflow. BigIndex also enables additional complex functionality, such as distributed indexing and faceted search, which in turn provides a high degree of flexibility in custom application development and search technology choices.

Adaptive MapReduce for job acceleration

Jobs running on BigInsights often 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 API operations.

Jaql for a big data query language

A powerful, high-level declarative query language developed by IBM and contributed to the open source community, Jaql provides the capability to process both structured and unstructured data. It has a SQL-like interface that makes it easy to learn for developers familiar with SQL languages and helps simplify integration with relational databases.

Explore and visualize data

InfoSphere BigInsights includes a broad palette of analytics tools and capabilities to help organizations build powerful, custom analytic applications that deliver results and insights tailored to specific business needs.

IBM BigSheets

A revolution in data analysis tools, BigSheets is a browser-based, spreadsheet-like tool enabling data scientists and business users to explore data stored in InfoSphere BigInsights applications and create analytic queries without writing any code. Built-in analytic macros address common data exploration requirements, further improving data accessibility.

Version 2.0 of InfoSphere BigInsights updates BigSheets to include chart customization features and browsing capabilities that make it easier than ever to explore, manipulate and analyze data. A new application for HBase enables business users to access HBase directly from the InfoSphere BigInsights web console and view results in BigSheets.

BigSheets can help business users perform the following tasks:

- Integrate large amounts of unstructured data from web-based repositories
- Collect a wide range of unstructured data stemming from user-defined seed URLs
- Extract and enrich data using text analytics
- Explore and visualize data in specific, user-defined contexts

Integrate big data into existing information architectures

InfoSphere BigInsights is built on top of and extends open-source Hadoop with the capabilities needed to make it an enterprise-grade platform. The platform's open, flexible architecture gives organizations the choice of using the IBM-provided Apache Hadoop distribution or other Hadoop distributions, such as Cloudera's Distribution, including Apache Hadoop (CDH). Existing CDH installations can take advantage of InfoSphere BigInsights for its enterprise-class features such as text analytics, developer tooling and user-friendly data manipulation and exploration.

Hardware requirements and operating system support

- **Intel x86 servers, 64-bit, with a minimum of 4 GB memory and 40 GB of disk storage**
- **Red Hat Enterprise Linux 5.3 and above, Red Hat Enterprise Linux 6.2 and above, SUSE Linux Enterprise 11 and above, SP2 and Power Linux**

By its nature, open source software does not include technical support, and it may come with legal terms and conditions that do not suit many organizations. In contrast, InfoSphere BigInsights Enterprise Edition is delivered with standard IBM software licensing and support agreements. Organizations can deploy it under familiar licensing terms that help minimize uncertainty and risk—with the confidence that they will be backed by 24x7 support offerings, education and a worldwide professional services organization.

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 so InfoSphere BigInsights includes high-speed connectors for IBM DB2® database software, the IBM PureData™ System family of data warehouse appliances, IBM Netezza® appliances, IBM InfoSphere Warehouse and IBM Smart Analytics System. These high-speed connectors help simplify and accelerate data manipulation tasks. Moreover, the IBM InfoSphere DataStage® tool includes a connector to allow BigInsights data to be leveraged within a DataStage ETL job.

InfoSphere BigInsights comes with a standard JDBC connector, making it possible for organizations to quickly integrate with a wide variety of data and information systems, including Oracle, Microsoft SQL Server, MySQL and Teradata.

IBM InfoSphere Data Explorer

InfoSphere BigInsights V2.0 includes a limited-use license for the bundled InfoSphere Data Explorer software, which enables organizations to discover, navigate and visualize vast amounts of structured and unstructured information within InfoSphere BigInsights.

InfoSphere Streams

InfoSphere BigInsights includes a limited-use license for InfoSphere Streams, which enables real-time, continuous analysis of data on the fly. InfoSphere Streams is an enterprise-class stream processing system that can be used to extract actionable insights from data as it arrives in the enterprise, while transforming data and ingesting it into InfoSphere BigInsights at high speeds. InfoSphere Streams allows organizations to capture and act on business data in real time—rapidly ingesting, analyzing and correlating information as it arrives—and fundamentally enhance the processing component. It includes a connector that allows end users to read and write to the InfoSphere BigInsights file system.

Cognos Business Intelligence

InfoSphere BigInsights includes a limited-use license for IBM Cognos® Business Intelligence, which enables business users to access and analyze the information they need to enhance 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.

For more information

To learn more about BigInsights Enterprise Edition, please contact your IBM sales representative or visit: **ibm.com**/software/data/infosphere/biginsights/enterprise.html

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
January 2013

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.

♻ Please Recycle