IBM InfoSphere BigInsights Enterprise Edition

Cost-effectively store, manage and gain insights from big data

Highlights

- Delivers big data analytics on an enterprise-ready platform
- Integrates with non-IBM solutions as well as popular IBM offerings such as IBM® Netezza® appliances, IBM DB2® database software, IBM InfoSphere® Streams and IBM InfoSphere DataStage®
- Supports structured, semi-structured and unstructured data for maximum flexibility

Many companies are seeing dramatic growth in the variety, velocity and volume of information being generated by their businesses. Organizations are struggling with how to manage vast and diverse quantities of both traditional structured data and semi-structured or unstructured data types—large, untapped data sets that define a new category of information: big data. Organizations see tremendous potential for deep insights that drive fast, clear and nuanced decision making, but they 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 create new solutions that cost-effectively turn large, complex volumes of data into insight. This enterprise-ready analytics platform combines Apache Hadoop with unique IBM innovations to deliver massive scale-out data processing and analysis with built-in resiliency and fault tolerance.

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 big data to the enterprise

InfoSphere BigInsights takes open-source Hadoop and adds the enterprise-class functionality and integration necessary to help meet critical business requirements. Organizations can run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure leverages Hadoop's 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.

InfoSphere BigInsights provides a thoroughly tested and integrated solution that combines the benefits of leading-edge technologies with mature, enterprise-ready features. Administrators start with a GUI-driven installation tool that enables them to get up and running quickly. The guided installation lets administrators 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, InfoSphere BigInsights delivers enterprise-class features that help streamline workload management and system administration (see Figure 1). For example, robust job management features give organizations fine-grained control of all aspects of their 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 extensive details on configuration, tasks, attempts and other critical information. In addition, InfoSphere BigInsights provides extensive administration features, including Hadoop Distributed File System (HDFS) and MapReduce administration, cluster and server management, the ability to view HDFS file content, and role-specific views.

Figure 1: Explore tasks and links to web console functionality with InfoSphere BigInsights.

Test-drive it now: Download InfoSphere BigInsights Basic Edition

For developers, partners and anyone who wants to try out the technology, IBM offers InfoSphere BigInsights Basic Edition, a no-cost, entry-level edition of InfoSphere BigInsights. BigInsights Basic Edition can be downloaded and run in the cloud with the provider of your choice. For more information, visit: ibm.com/software/data/infosphere/biginsights/basic.html

Users of the basic edition have the option to purchase 24×7 Elite Support. Solutions developed using InfoSphere BigInsights Basic Edition can be seamlessly deployed to InfoSphere BigInsights Enterprise Edition.

Add security to big data analysis

Enterprises have very stringent requirements when it comes to security. 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 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.

Authorization

InfoSphere BigInsights provides four levels of user authorization, known as roles: System Administrator, Data Administrator, Application Administrator and Non-Administrative User. A user's access to data and features depends on the role that user is assigned.

Maximize performance, streamline job handling

IBM has added several capabilities that help increase performance and make InfoSphere BigInsights flexible and compatible with an enterprise environment.

BigInsights Scheduler for workflow allocation

Not all workloads have the same priority, and 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 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 BigInsights. Based on Apache Lucene, 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 InfoSphere BigInsights often end up creating multiple small tasks that consume a disproportionately large amount of system resources. To combat this, IBM has invented a technique called Adaptive MapReduce that speeds 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.

Dive deep into big data with analytics accelerators

InfoSphere BigInsights includes a broad palette of analytics tools and capabilities. 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.

BigSheets

Representing a revolution in data analysis tools, BigSheets is a browser-based tool enabling business users to explore data stored in BigInsights clusters and create analytic queries

without writing any code (see Figure 2). Built-in analytic macros address common data exploration requirements, further improving data accessibility. BigSheets can help business users:

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

BigSheets brings big data analytics to enterprise business users, giving them the tools to perform ad hoc analytics—and develop their own insights—without requiring IT support.

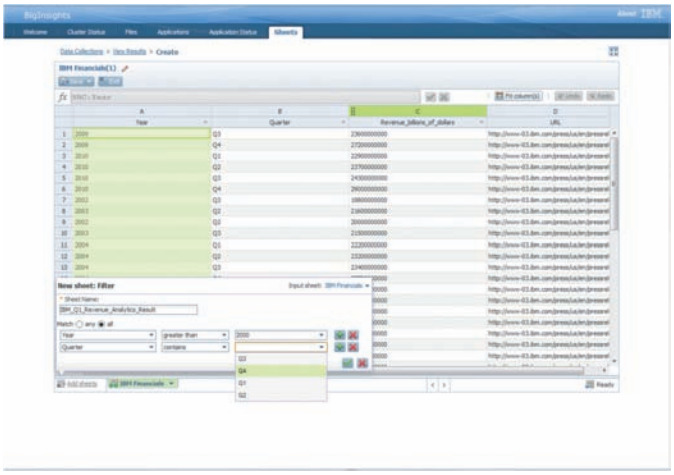

Figure 2: The BigSheets browser-based tool offers fast access to unstructured data and the ability to quickly create queries without writing code.

Advanced text analytics accelerator

BigInsights includes the powerful text analytics engine developed by IBM and used by IBM Watson™ on the *Jeopardy!* quiz show to defeat two of the game's best players. 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. The text analytics engine supports English, Dutch/Flemish, French, German, Italian, Portuguese, Spanish, Japanese and Chinese.

Jaql

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. InfoSphere BigInsights comes with prebuilt Jaql modules: IBM includes Jaql modules for Lucene indexes, the IBM Netezza family of data warehouse appliances, HBase (the Hadoop database) and workflows including the built-in text analytics capabilities of InfoSphere BigInsights.

Integrate big data into existing information architectures

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 Enterprise Edition includes high-speed connectors for the IBM Netezza family of data warehouse appliances, IBM DB2, IBM InfoSphere Warehouse and IBM Smart Analytics System. When used with DB2 or IBM warehouse offerings, these high-speed connectors help simplify and accelerate data manipulation tasks.

InfoSphere BigInsights Enterprise Edition also comes with a standard Java Database Connectivity (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.

The InfoSphere DataStage tool includes a connector to allow BigInsights data to be leveraged within a DataStage ETL job. IBM InfoSphere Streams includes a connector that allows end users to read and write to the BigInsights file system.

InfoSphere BigInsights also supports open development standards and the Apache Nutch web-search software for crawling unstructured data content both inside and outside the enterprise.

Obtain enterprise-class support

By its nature, open source software does not include technical support, and it may come with legal terms and conditions that do not suit some 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 24×7 support offerings, education and a worldwide professional services organization.

IBM brings open source technologies to organizations in solutions that are comprehensive, integrated and enterprise-ready. IBM helps companies simplify and accelerate the introduction of big data technologies by integrating them into their existing information supply chains. Also, organizations know that IBM can deliver the performance and reliability that they need, in a solution that will become an integral part of their decision-making processes.

Hardware requirements and operating system support

Intel x86 servers, 64-bit, with a minimum of 4 GB memory and 30 GB of disk storage

Red Hat Enterprise Linux 5.3, 5.4, 5.5, 5.6, 6.0

SUSE Linux Enterprise 11, 11 SP1

Note: Refer to product documentation for up-to-date operating system version support: ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&htmlfid=897/ENUS211-501&appname=USN

For more information

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

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589 U.S.A.

Produced in the United States of America
January 2012

Please Recycle

IMD14385-USEN-00