IBM InfoSphere BigInsights Enterprise Edition

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 popular IBM offerings as well as third-party solutions

Tame big data

IBM® InfoSphere® BigInsights™ Enterprise Edition 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.

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 poly-structured data without straining the data warehouse. As a preprocessing hub—also referred to a "landing zone" for data—InfoSphere 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 InfoSphere BigInsights, analytics is not a separate step performed after data is stored; instead, InfoSphere BigInsights, in combination with InfoSphere Streams, enables real-time analytics that can leverage historic models derived from data being analyzed at rest. InfoSphere 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. 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 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 InfoSphere 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

InfoSphere BigInsights 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. InfoSphere 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 through Big SQL. A SQL interface over Hadoop that is built on established standards, Big SQL leverages standard SQL to allow users to access big data in the same way they leverage other relational data. InfoSphere 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 in InfoSphere BigInsights

InfoSphere 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. Once the raw data has been stored across the distributed cluster, the systems can efficiently handle queries and data analysis.

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 InfoSphere 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, InfoSphere BigInsights provides administration features for Hadoop Distributed File System (HDFS), IBM General Parallel File System (GPFS™) File Placement Optimizer (FPO), big data applications and MapReduce jobs, and cluster management.

Try InfoSphere BigInsights at no cost

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

As shown in Figure 1, InfoSphere BigInsights provides several enterprise capabilities. The next sections will walk through each area.

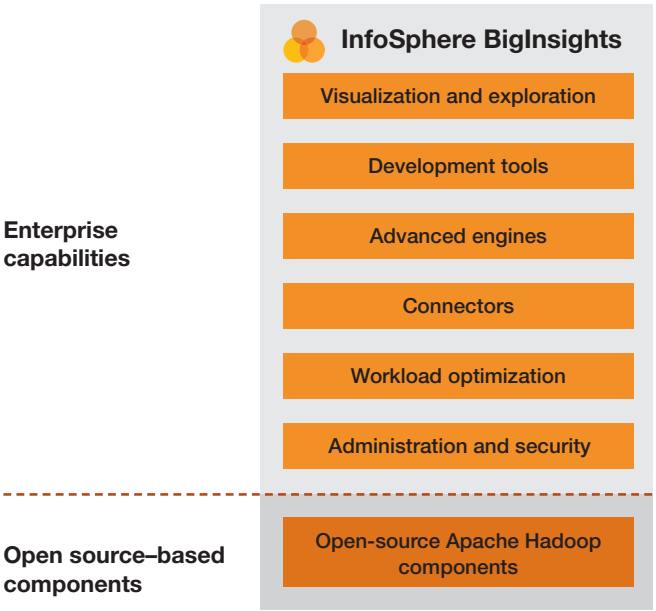

Figure 1. InfoSphere BigInsights adds enterprise capabilities to open-source components.

Visualization and exploration

InfoSphere 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.

IBM BigSheets, data exploration and dashboards

IBM BigSheets is a browser-based tool that enables data scientists and business users to explore data stored in Hadoop via a spreadsheet-style interaction model. Built-in analytic macros address common information discovery needs. Chart customization features and pivot table capabilities in BigSheets make it easier than ever to explore, manipulate and analyze big data.

BigSheets can help business users perform the following tasks:

- Integrate and explore large amounts of poly-structured data
- Extract and enrich data using text analytics
- Explore and visualize data in specific, user-defined contexts

InfoSphere 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 Hadoop/InfoSphere BigInsights cluster.

Development tools

InfoSphere 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.

InfoSphere 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—all from within the InfoSphere BigInsights Eclipse tooling.

Advanced engines and accelerators

InfoSphere 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

InfoSphere BigInsights includes a powerful text analytics engine developed by IBM. Using a comprehensive library of rules or by developing their own custom rules, BigInsights 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

Introduced in InfoSphere BigInsights V2.0, 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 and business intelligence including sentiment, purchase intent and product/service ownership as well as micro-segmentation 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

Also introduced 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. 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 InfoSphere 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 BigInsights Hadoop file system, while InfoSphere BigInsights can directly ingest unstructured data.

InfoSphere BigInsights also features 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. Moreover, the IBM InfoSphere DataStage® tool includes a connector that enables InfoSphere BigInsights data to be leveraged within an InfoSphere DataStage extract/transform/load (ETL) job. 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.

Workload optimization

InfoSphere BigInsights provides several features that help increase performance, as well as enhance its adaptability and compatibility within 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, 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. 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 Pluggable Authentication Modules (PAM) 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. And integration of InfoSphere 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.

Enhanced enterprise integration

IBM InfoSphere Data Explorer

InfoSphere BigInsights includes a limited-use license for InfoSphere Data 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

InfoSphere 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 InfoSphere 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

InfoSphere 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.

Increased consumability and functionality

Version 2.1 of InfoSphere BigInsights includes several new features that are designed for greater ease of use.

Big SQL: A new SQL interface for InfoSphere BigInsights

New in InfoSphere BigInsights V2.1 is Big SQL, a native SQL query engine that enables SQL access to data stored in InfoSphere BigInsights (Figure 2). Big SQL comes with a standards-compliant JDBC and Open Database Connectivity (ODBC) driver, and supports a rich SQL syntax. This allows new users to immediately apply their existing SQL skills in a big data environment, querying data in Hadoop just as they query data from their database applications. With Big SQL, users can leverage MapReduce parallelism for complex data sets and bypass it for low latency in small queries (for example, sub-second HBase queries).

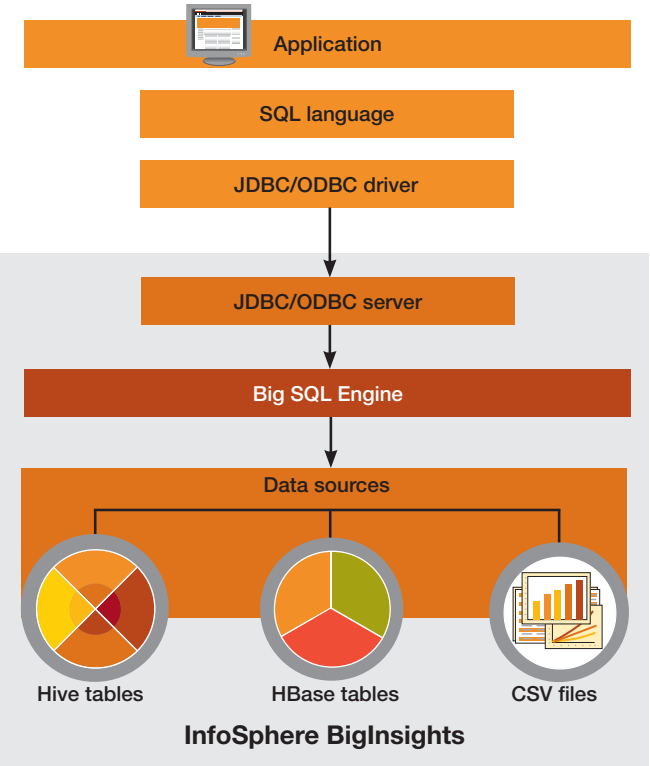

Figure 2. The new Big SQL query engine enables access to data stored in InfoSphere BigInsights.

Cognos BI: Optimized for Big SQL

Big SQL enables the Cognos BI server to push many types of computations to InfoSphere BigInsights MapReduce processing instead of computing them locally at the cost of performance as it would do with Hive. Processing queries closer to the data results in faster response times and frees Cognos from the latency and limitations of querying Hadoop via Hive.

GPFS FPO

Another new feature in InfoSphere BigInsights V2.1 is GPFS FPO, which is an enterprise-grade Portable Operating System Interface (POSIX)-compliant file system that provides an alternative to HDFS. Based on GPFS and built nearly 12 years ago for high-performance computing (HPC) functions, this highly regarded file system brings the value of a proven distributed file system with integrated information lifecycle management to big data environments.

The GPFS FPO distributed metadata feature eliminates single points of failure that can bring your analytics capabilities to a halt. With full POSIX compliance support, and the ability to work with traditional applications that require read/write capabilities, you can leverage GPFS FPO for all of your file-based access requirements. It also addresses a rising concern around big data security by offering tighter access control.

High availability

InfoSphere BigInsights V2.1 delivers out-of-the-box high availability with seamless, automatic and transparent failover. The central gatekeeper for running HDFS operations and metadata management for HDFS clients and DataNodes is the HDFS NameNode. It maintains HDFS metadata for file and directory listing, security, file and data blocs mapping locations to facilitate data transfer between the HDFS client and DataNodes. When there is a NameNode failure, the NameNode process becomes the single point of failure for the cluster and requires administrative intervention, which can yield long downtimes.

The NameNode High Availability solution in InfoSphere BigInsights V2.1 detects NameNode failures and performs automatic failover to a standby node, eliminating the need for administrator intervention and drastically improving system availability. Moreover, the failover is seamless to HDFS clients and DataNodes.

Accelerators and BigSheets enhancements

InfoSphere BigInsights V2.1 features upgrades to two previously-introduced accelerators: Machine Data Analytics Accelerator and Social Data Analytics Accelerator. A new configuration user interface in the Machine Data Accelerator provides an easy and intuitive way to perform workflow configuration, significantly enhancing time-to-value. The Machine Data Accelerator also now supports more data sources and analyzes InfoSphere BigInsights/Hadoop logs as well. Improvements to the Social Data Accelerator significantly advance scalability and performance. Users can now add dictionary terms dynamically to enhance their analysis results.

InfoSphere BigInsights V2.1 also enhances data discovery and visualization capabilities by embedding text analytics capabilities directly into BigSheets. Many InfoSphere BigInsights text analytics extractors are available via built-in functions, allowing users to extract names, addresses, organizations, emails, phone numbers and so on from within BigSheets.

Enterprise-class support and service

By its nature, open source software does not include technical support. In contrast, InfoSphere BigInsights Enterprise Edition is delivered with standard IBM software 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.

Hardware requirements and operating system support:

- **x86 or IBM Power® 64-bit systems with a minimum of 8 GB memory and 40 GB of disk storage**
- **Red Hat Enterprise Linux 5.6 and above, Red Hat Enterprise Linux 6.1 and above, SUSE Linux Enterprise 11 SP 2 and above, SP2 and IBM PowerLinux™**

For more information

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

© Copyright IBM Corporation 2013

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
August 2013

IBM, the IBM logo, ibm.com, BigInsights, Cognos, DataStage, DB2, GPFS, Guardium, InfoSphere, Power, PowerLinux, and PureData are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at **ibm.com**/legal/copytrade.shtml

Netezza is a trademark or registered trademark of IBM International Group B.V., an IBM Company.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft is a trademark of Microsoft Corporation in the United States, other countries, or both.

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.

Actual available storage capacity may be reported for both uncompressed and compressed data and will vary and may be less than stated.

Please Recycle

IMD14385-USEN-02