Understanding InfoSphere BigInsights

An introduction for software architects and technical leaders

Skill Level: Introductory

Cynthia M. Saracco (saracco@us.ibm.com)
Senior Solution Architect
IBM

06 Oct 2011

Perhaps you've heard about InfoSphere® BigInsights, IBM®'s software platform for storing and analyzing “big data,” and you may be wondering what all the buzz is about. This article provides an introduction to BigInsights and explains what the product was designed to do, when it can be useful, and how it can complement other software you may already have.

Architectural overview

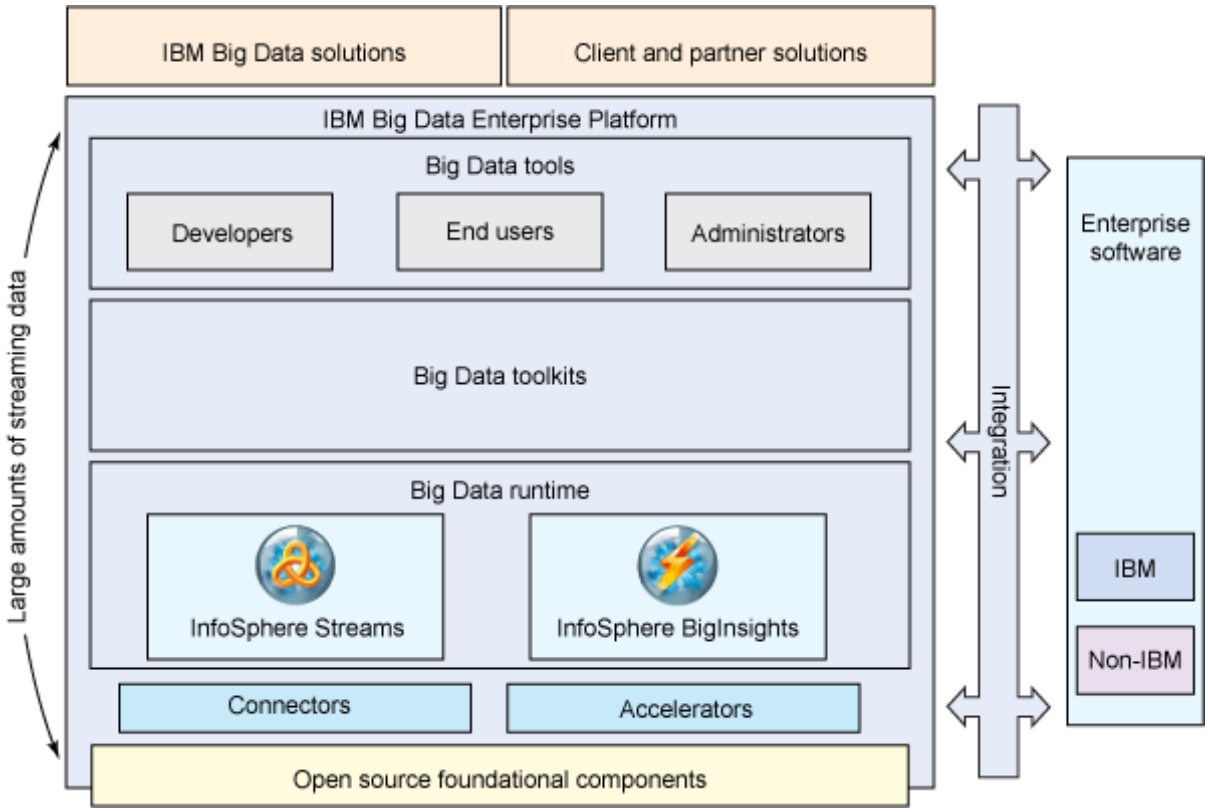
BigData University
Develop skills working with Hadoop and big data analytics at Big Data University.

InfoSphere BigInsights 1.2 is a software platform designed to help firms discover and analyze business insights hidden in large volumes of a diverse range of data—data that's often ignored or discarded because it's too impractical or difficult to process using traditional means. Examples of such data include log records, click streams, social media data, news feeds, electronic sensor output, and even some transactional data. To help firms derive value from such data in an efficient manner, BigInsights incorporates several open source projects (including Apache™ Hadoop™) and a number of IBM-developed technologies.

Figure 1 illustrates IBM's big data platform, which includes software for processing streaming data and persistent data. BigInsights supports the latter, while InfoSphere

Streams supports the former. The two can be deployed together to support real-time and batch analytics of various forms of raw data, or they can be deployed individually to meet specific application objectives. The remainder of this article focuses on BigInsights. For more about InfoSphere Streams, see the Resources section.

Figure 1. IBM's platform and vision for big data



IBM developed BigInsights to help firms process and analyze the increasing volume, variety, and velocity of data of interest to many enterprises. Consider that industry analysts expect the quantity of digital data to increase rapidly in the coming years. Indeed, one firm, International Data Corporation, expects volumes to grow up to 44 times by 2020 when compared with 2009 levels, and most of that data will be in unstructured or semi-structured formats. As a result, many IT professionals anticipate new data processing challenges; they often use the term "Big Data" (or "big data") to refer to this issue.

However, many firms recognize that analyzing large volumes of raw data can reveal patterns and insights important to their organizations. Application areas span many domains, including customer retention, customer service, market intelligence, business planning and operations, scientific research, security, and other areas. Such applications may require analyzing application, system, or sensor log data; consumer or public sentiment expressed through various electronic venues; text-rich data, including documents, emails, and messages; and various other sources of

data. Unfortunately, the sheer effort involved to collect, process, analyze, and manage this data can seem daunting.

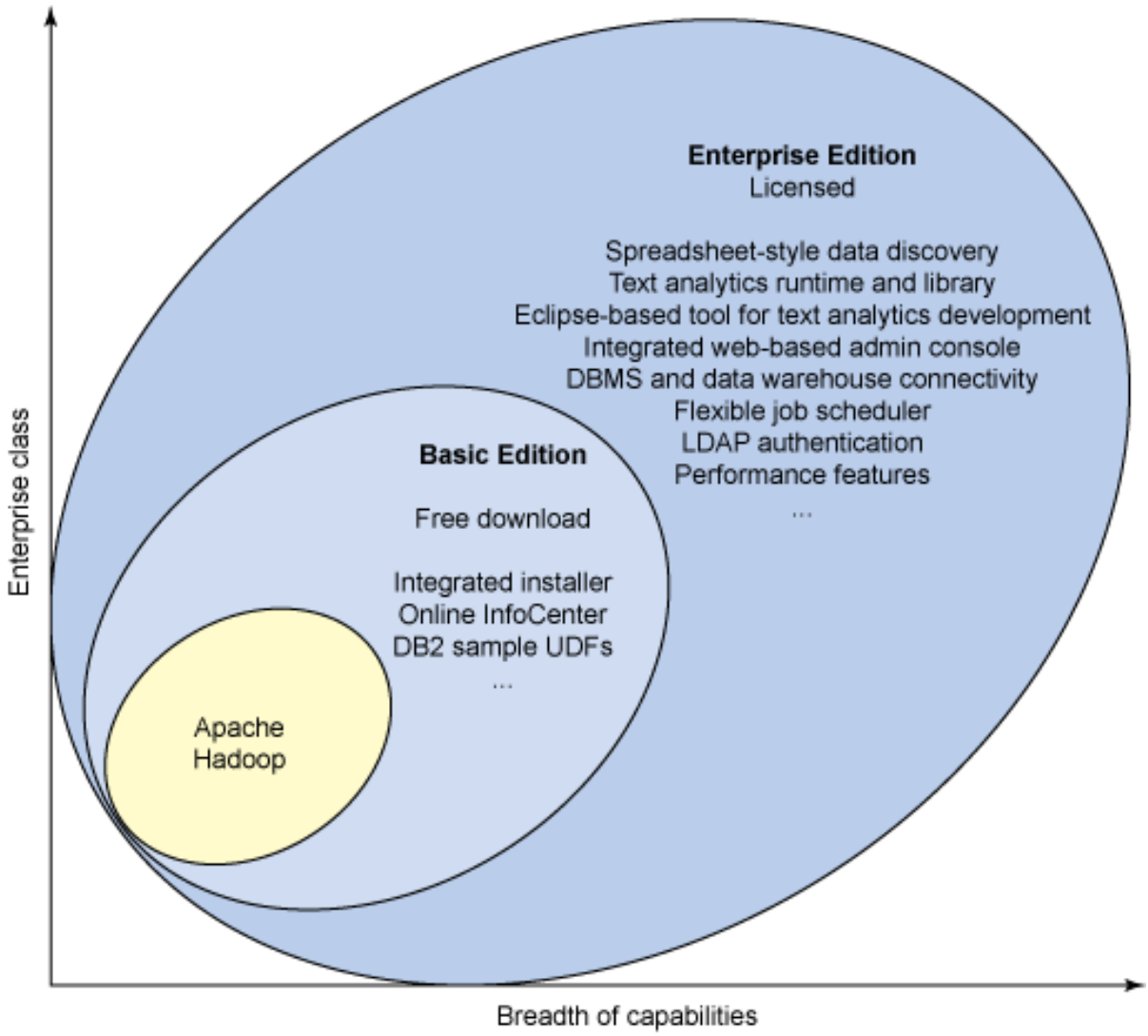
To make sifting through large volumes of diverse data practical, BigInsights provides built-in analytic technologies and exploits shared-nothing hardware clusters. It transparently distributes data stored in files across disks attached to various nodes in a cluster, directing subtasks of applications to processors that are close to the target subsets of your data. This approach minimizes network traffic and improves runtime performance. For fault tolerance, BigInsights automatically replicates each portion of your data on multiple disks based on parameters specified by an administrator. Such replication enables BigInsights to automatically recover from a disk or node failure by redirecting work elsewhere.

BigInsights doesn't replace a relational database management system (DBMS) or a traditional data warehouse. It isn't optimized for interactive queries over tabular data structures, online analytical processing (OLAP), or online transaction processing (OLTP) applications. Rather, it's a platform that can augment your existing analytic infrastructure, enabling you to filter through high volumes of raw data and combine the results with structured data stored in your DBMS or warehouse, if desired. Potential integration scenarios will be discussed later.

Basic and Enterprise Editions

By now, you may be wondering about the technologies that comprise the BigInsights platform. A number of IBM and open source technologies are part of BigInsights, which is available in two editions: Basic and Enterprise. As shown in Figure 2, both editions include Apache Hadoop and other open source software, which are explained in more detail later.

Figure 2. InfoSphere BigInsights 1.2 Basic and Enterprise Editions



Basic Edition is available for free download and can manage up to 10 TB of data. As such, it's suitable for pilot projects and exploratory work. The Enterprise Edition is a fee-based offering with no licensing restrictions on the quantity of data that can be managed. It includes all the features of the Basic Edition and offers additional analytic, administrative, and software integration capabilities. As such, Enterprise Edition is suitable for production applications.

Open source technologies

You may already be familiar with certain open source projects, so explore these first. Open source projects included with BigInsights 1.2 Basic and Enterprise Editions are:

- **Apache Hadoop** (including the Hadoop Distributed File System (HDFS),

MapReduce framework, and common utilities), a software framework for data-intensive applications that exploit distributed computing environments

- **Pig,** a high-level programming language and runtime environment for Hadoop
- **Jaql,** a high-level query language based on JavaScript Object Notation (JSON), which also supports SQL.
- **Hive,** a data warehouse infrastructure designed to support batch queries and analysis of files managed by Hadoop
- **HBase,** a column-oriented data storage environment designed to support large, sparsely populated tables in Hadoop
- **Flume,** a facility for collecting and loading data into Hadoop
- **Lucene,** text search and indexing technology
- **Avro,** data serialization technology
- **ZooKeeper,** a coordination service for distributed applications
- **Oozie,** workflow/job orchestration technology

These projects are well-documented at publicly accessible websites. See the Resources section for links to introductory materials on Hadoop and related projects.

IBM technologies

In addition to open source software, BigInsights includes a number of IBM-developed technologies to help you become productive quickly. Examples include a text analysis engine and supporting development tool, a data exploration tool for business analysts, enterprise software integration, and various platform enhancements to simplify administration and help improve runtime performance. Take a closer look.

Text-based analytics and tooling

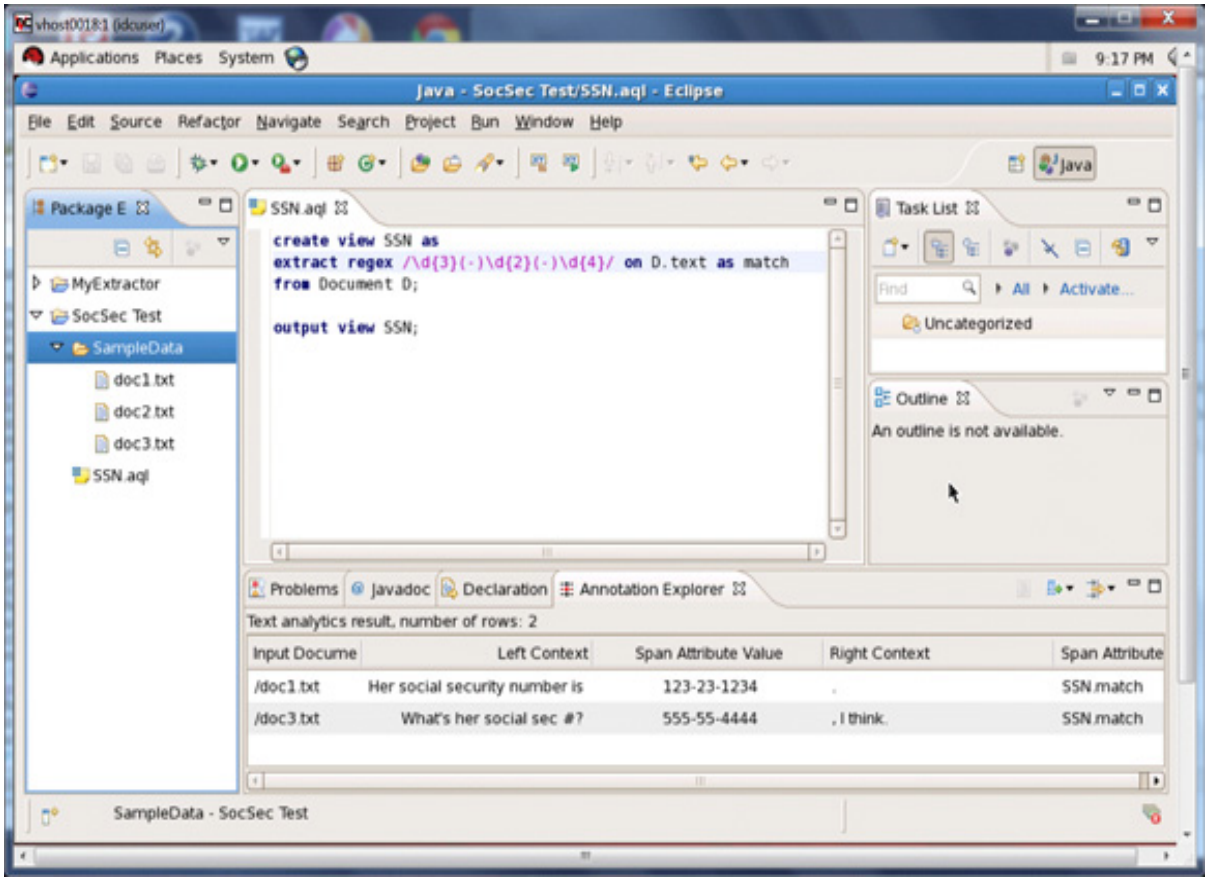
As previously mentioned, BigInsights is designed to help firms analyze a diverse range of data, including data that's loosely structured or largely unstructured. Various types of text data fall into this category. Indeed, financial documents, legal documents, marketing collateral, emails, blogs, news reports, press releases, and social media websites contain text-based data that firms may want to process and assess.

For this reason, BigInsights Enterprise Edition includes a text processing engine and library of annotators that enable developers to query and identify items of interest in

documents and messages. Examples of business entities that BigInsights can extract from text-based data include persons, email addresses, street addresses, phone numbers, URLs, joint ventures, alliances, and others.

In addition, programmers can use the Eclipse-based plug-in to build their own library of text analytic functions for BigInsights. Shown in Figure 3, the plug-in includes an expression builder, pattern discovery technology, a test environment, and a results explorer to promote rapid prototyping and refinement of complex text analytic functions tailored to specific application requirements.

Figure 3. BigInsights includes an Eclipse plug-in for creating new text analytic functions (or annotators)



In this figure, the central box displays the syntax for creating a "SSN" annotator to identify USA social security numbers. The annotator's definition calls for 3 single-digit numbers, a hyphen, 2 single-digit numbers, a hyphen, and 4 single-digit numbers. Although this example is extremely simple, it illustrates the basic concept of how to define an annotator. Of course, programmers can use this tool to build complex, production-quality annotators, including annotators that build on previously-created annotators.

Referring back to the figure again, you'll see that the pane on the left indicates that

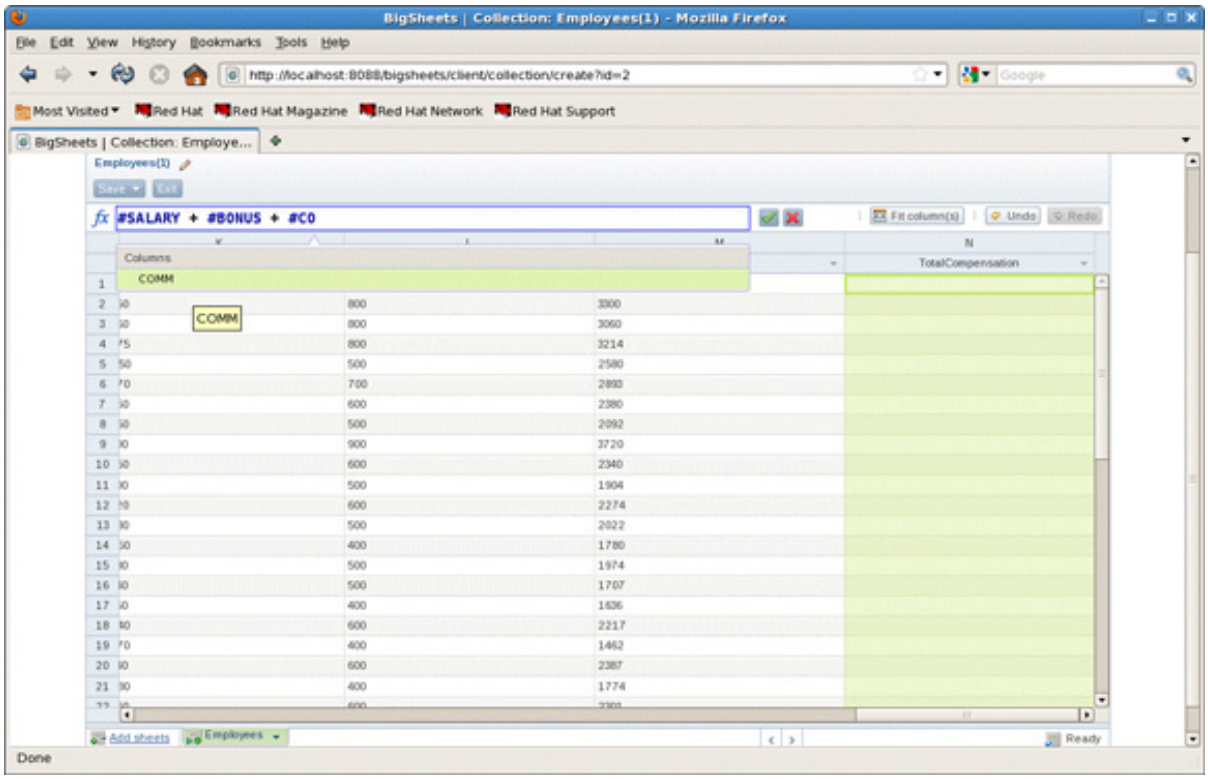
three sample documents that were imported into the project for test purposes. The bottom pane illustrates the result of a test run using these documents as input. The results show the "SSN" entities that the annotator identified, the context in which each SSN appeared in the document (that is, the text immediately to the left and right of the identified social security number), and the name of the source document. Once the annotator (or a set of annotators) is completed, a developer can export the resulting code in the form of an Annotation Operator Graph (AOG) for use by BigInsights applications.

Spreadsheet-like data discovery and exploration

To help business analysts and non-programmers work with "big data," BigInsights Enterprise Edition provides a spreadsheet-like data analysis tool. Launched through a Web browser, BigSheets enables business analysts to create *collections* of data to explore. To create a collection, an analyst specifies the desired data source(s), which might include the BigInsights distributed file system, a local file system, or the output of a Web crawl. BigSheets provides built-in support for popular data formats, such as JSON data, comma-separated values (CSV), tab-separated values (TSV), character-delimited data, and others. If desired, programmers can create plug-ins to process additional data formats and execute custom functions.

When an analyst executes (or runs) the collection's definition, BigSheets generates MapReduce jobs behind the scenes to retrieve and process the necessary data. Analysts can also review and manipulate the collection's data using built-in functions and macros. Such work is done through a traditional spreadsheet-like interface, as shown in Figure 4 which depicts a simple user-defined formula for populating a new column with values derived from other columns in the collection.

Figure 4. BigInsights Enterprise includes a spreadsheet-based analytic tool



Finally, analysts can use charting facilities in BigSheets to visualize some or all of their collection's content, if desired. In addition, they can export collection data in one of several popular formats for use by other applications. HTML, CSV, and JSON are some of the supported export formats.

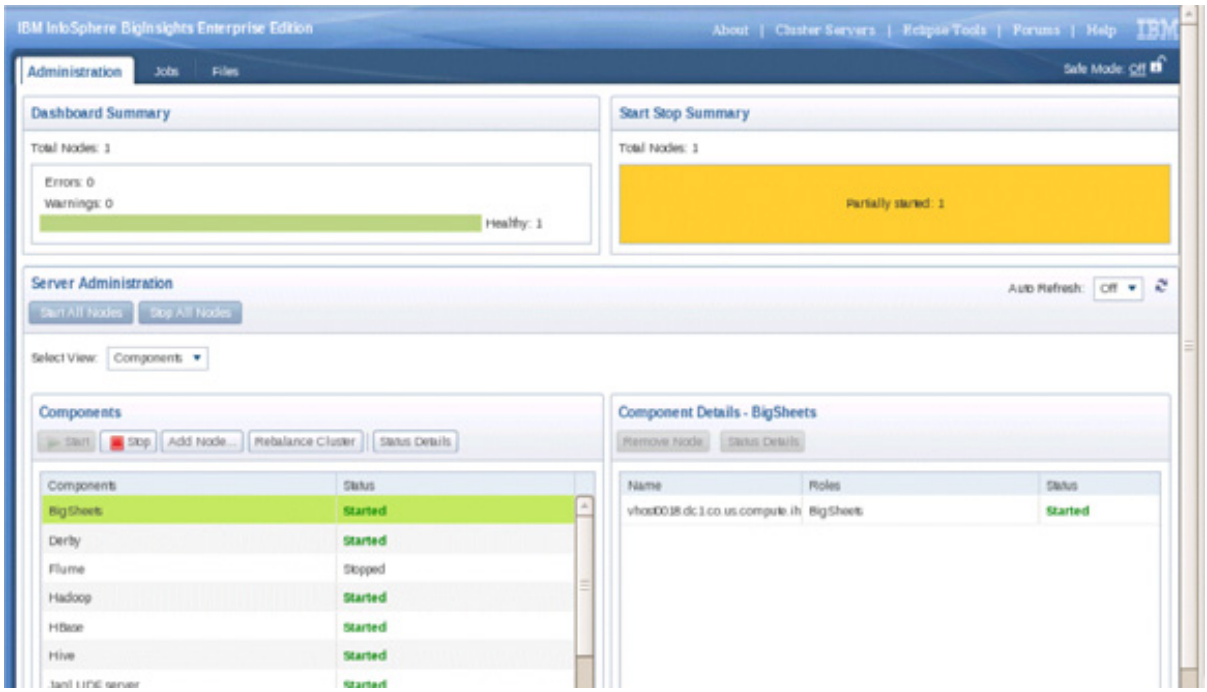
Integrated installation and administration tools

To help firms get off to a quick start, BigInsights Basic and Enterprise editions provide a Web-based tool that installs and configures all supported IBM and non-IBM software selected by an administrator. Details about the progress of a BigInsights installation are reported in real time, and a built-in "health check" mechanism automatically verifies and reports on the success of the installation.

By contrast, those working with individual open source offerings would need to iteratively download, configure, and test each software project they wanted to use. Furthermore, they would need to be sensitive to any software pre-requisites and incompatibilities that might exist among the desired projects.

Once BigInsights is installed, Enterprise administrators can work with a Web-based management console to inspect the status of their BigInsights environment at any time. Through this console, they can start and stop nodes, investigate the status of MapReduce jobs, review log records, assess the overall health of the system, start and stop optional components, navigate the distributed file system, and more. Figure 5 illustrates a portion of the primary panel of the Web console.

Figure 5. A portion of the BigInsights Enterprise Web console



Enterprise software integration

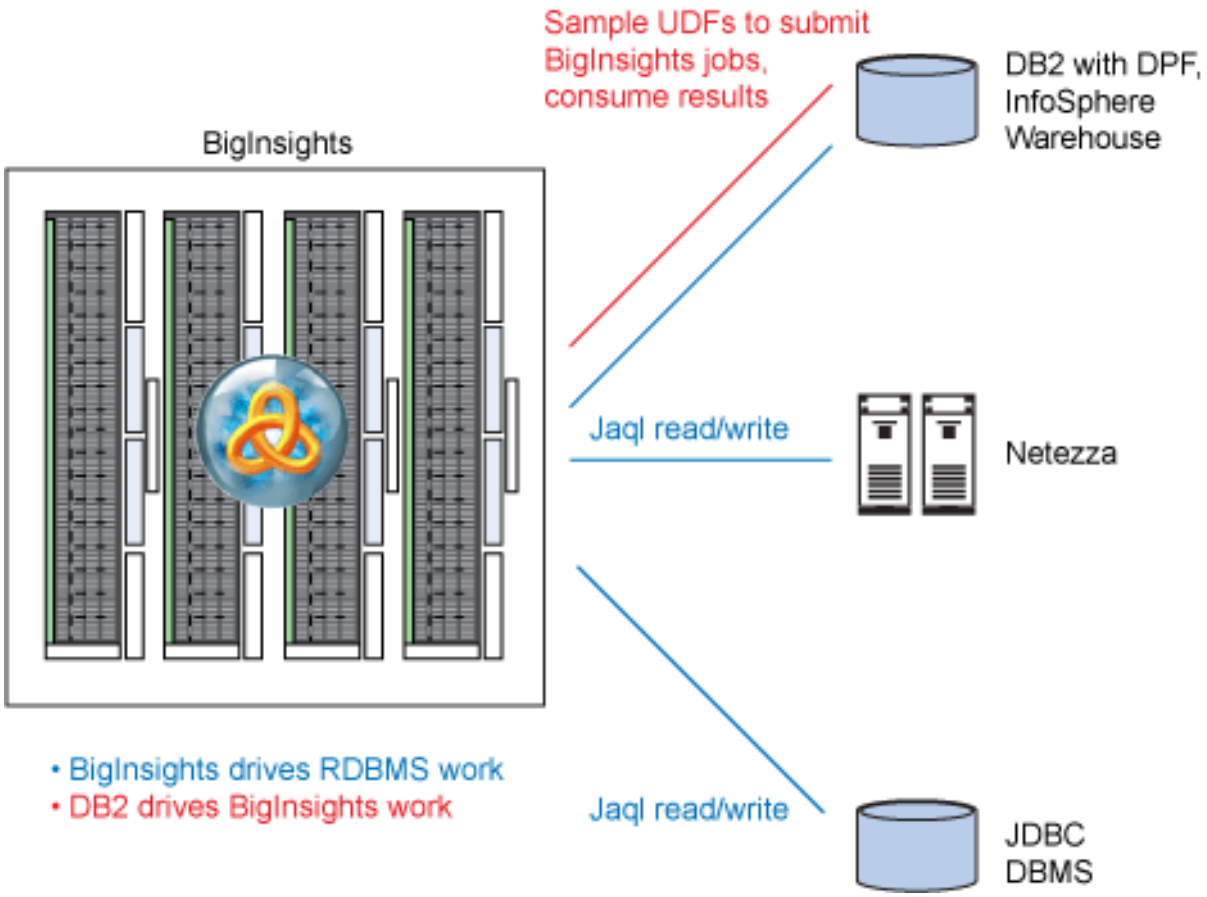
Many organizations are concerned about introducing yet another information management platform into their existing IT infrastructure. Quite commonly, IT architects worry about integrating information managed by the new system with other important data they already have in their enterprise.

To address this concern, BigInsights Enterprise Edition provides Jaql developers with JDBC connectivity to Netezza and DB2 so they can transfer data to and from these sources in a manner that exploits the native parallel processing capabilities of those platforms. Such support is useful for BigInsights developers who want to join reference data stored in a relational DBMS with data managed by BigInsights. To access other relational data sources, BigInsights provides a generic JDBC connector.

In addition, both BigInsights Basic and Enterprise Editions provide sample DB2 user-defined functions (UDFs) that allow DB2 programmers to launch Jaql queries in BigInsights, join the output with DB2 data, and present the results to DB2 users and applications. These UDFs can be registered with a DB2 9.7 server for Linux, Unix, and Windows platforms.

Figure 6 illustrates the DBMS and data warehouse connectivity provided through BigInsights 1.2.

Figure 6. DBMS and data warehouse connectivity for BigInsights 1.2



Platform enhancements and performance features

While BigInsights uses open source technologies that offer strong runtime performance and high levels of scalability, the Enterprise Edition also employs IBM-specific software to further enhance administration and performance.

For example, BigInsights offers an optional job scheduling mechanism for fine tuning resource allocation among long-running and short-running jobs. Administrators can use a property setting to allocate maximum resources to small jobs to help ensure they complete quickly. This job scheduling option is available in addition to Hadoop's first in/first out (FIFO) and "fair" scheduling approaches.

In addition, BigInsights provides enhanced security by supporting LDAP authentication to its Web console. LDAP and reverse proxy support help administrators restrict access to users with appropriate authorization.

Performance enhancements include efficient processing of text-based compressed data through the use of IBM LZO-based compression technology. BigInsights also includes adaptive runtime techniques for Jaql jobs that can help improve runtime performance of target applications. When IBM's adaptive MapReduce technology is turned on (through a property setting or a Jaql option), Map tasks communicate

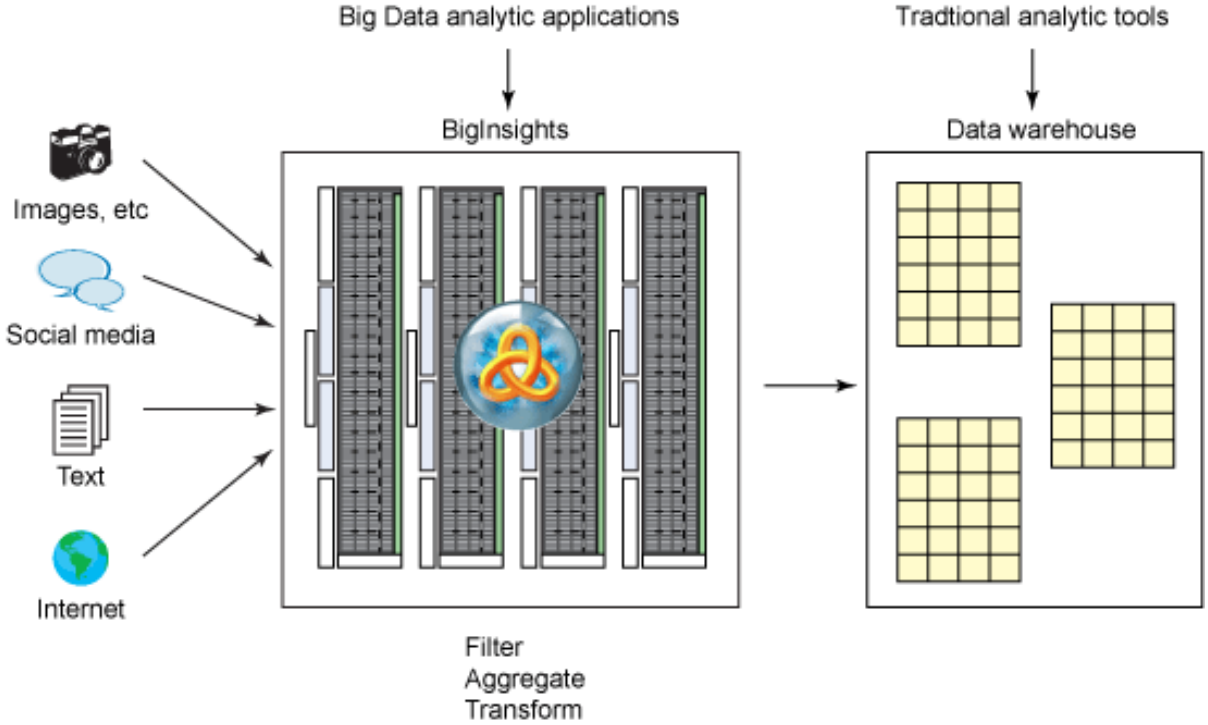
through ZooKeeper to understand the global state of the job. When it makes sense to do so, Map tasks can dynamically take on additional work, which can lead to improved runtime performance for the overall job.

How BigInsights fits into an enterprise data architecture

Working with big data is becoming an integral part of the enterprise data strategy at many firms. Indeed, a number of organizations are looking to deploy a software platform such as BigInsights so that they can manage big data from the moment it enters their enterprise. After storing the raw data in BigInsights, firms can manipulate, analyze, and summarize the data to gain new insights as well as feed downstream systems. In this manner, both the original (raw) data and modified forms are accessible for further processing.

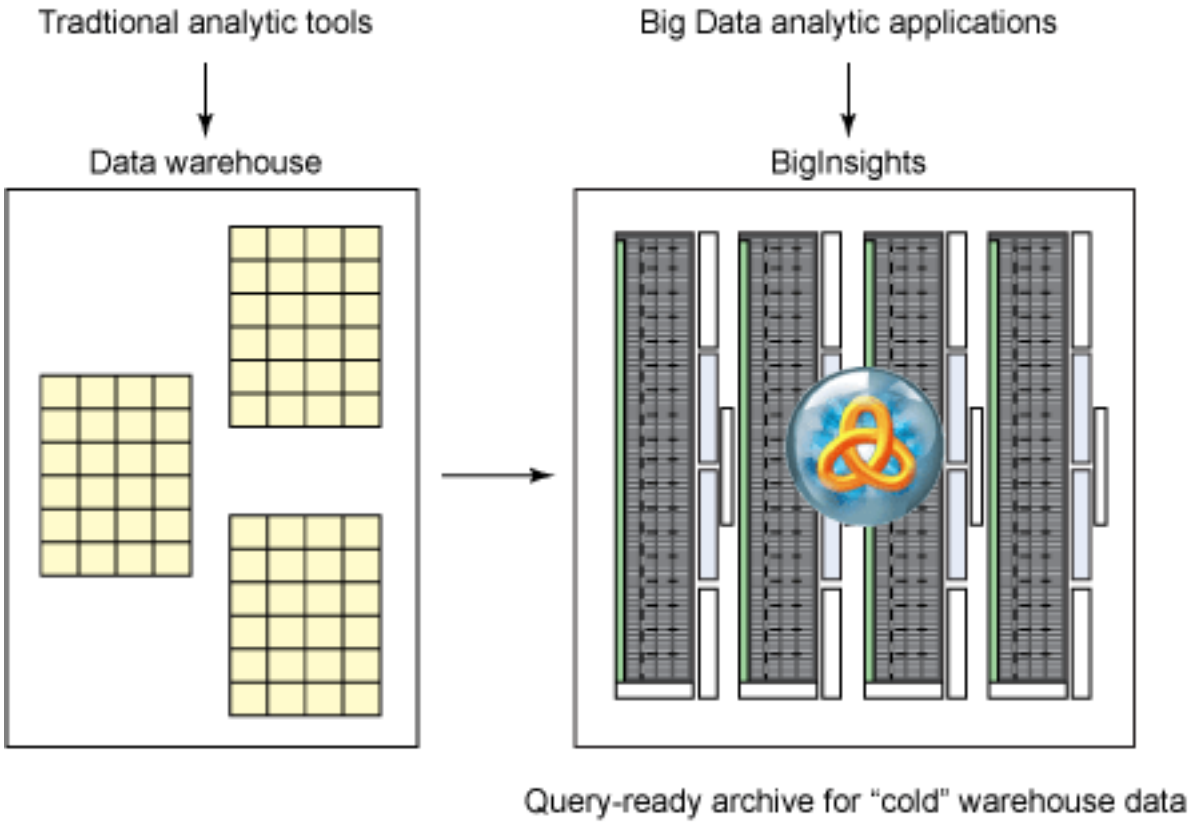
One potential deployment approach involves using BigInsights as a source for a data warehouse. BigInsights can sift through large volumes of unstructured or semi-structured data, capturing relevant information that can augment existing corporate data in a warehouse. Figure 7 illustrates such a scenario, which offers firms the ability to broaden their analytic coverage without creating an undue burden for their existing systems. Once in the warehouse, traditional business intelligence and query/report writing tools can work with the extracted, aggregated, and transformed portions of raw data stored in BigInsights.

Figure 7. Using BigInsights to filter and summarize big data for the warehouse



Another potential deployment approach involves using BigInsights as a query-ready archive for a data warehouse. With this approach, illustrated in Figure 8, frequently accessed data can be maintained in the warehouse while “cold” or outdated information can be offloaded to BigInsights. This allows firms to manage the size of their existing data management platforms while servicing the well-established needs of their existing applications. Offloading rarely queried data to BigInsights allows that data to remain accessible to applications that may have an occasional or unpredictable need to work with it.

Figure 8. BigInsights serving as a query-ready archive for a data warehouse



Summary

Helping firms manage, analyze and benefit from big data is a key area of focus for IBM. In this article, you were introduced to InfoSphere BigInsights, IBM's software platform for storing and analyzing such data. Based on open source and IBM-developed technologies, BigInsights is available in two editions: a free, Basic Edition suitable for exploratory projects and an Enterprise Edition suitable for production applications.

These are still the early days of big data, but not too early to get started leveraging it in a context that makes sense for your business. Analysts and early adopters generally agree that capitalizing on big data is an important information management

initiative. If you're ready to get started, consult the Resources section for links to free training materials and software.

Resources

Learn

- Visit the BigInsights Technical Enablement Wiki for links to technical materials, demos, training courses, news items, and more.
- Visit IBM's big data Web site to learn more about its big data platform and offerings. You may also want to visit the InfoSphere Streams Web site for details on how IBM's platform supports streaming data.
- Refer to the BigInsights InfoCenter for online documentation about the product.
- Stay tuned for a Flashbook on *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* to become available through the IBM Bookstore in late 2011.
- Read a summary of a forecast from International Data Corporation (IDC) about the growth of big data.
- Visit the Apache Hadoop Web site for details about Hadoop and related open source projects.
- Check out BigData University for free courses on big data and Hadoop.
- Learn more about Information Management at the developerWorks Information Management zone. Find technical documentation, how-to articles, education, downloads, product information, and more.
- Stay current with developerWorks technical events and webcasts.
- Follow developerWorks on Twitter.

Get products and technologies

- Build your next development project with IBM trial software, available for download directly from developerWorks.

Discuss

- Check out the developerWorks blogs and get involved in the developerWorks community.

About the author

Cynthia M. Saracco

Cynthia M. Saracco is a senior solutions architect at IBM's Silicon Valley Laboratory who specializes in emerging technologies and information management. She has 25 years of software industry experience, has written 3 books and more than 70 technical papers, and holds 7 patents.