Better Business Decisions at a Lower Cost with IBM InfoSphere BigInsights

IBM Redbooks Solution Guide

As activities in our world become more integrated, the rate of data growth is increasing exponentially. This data explosion (Figure 1) is referred to as *big data*, which renders current data management methods inadequate. IBM® is preparing the next generation of technology to meet these data management challenges.

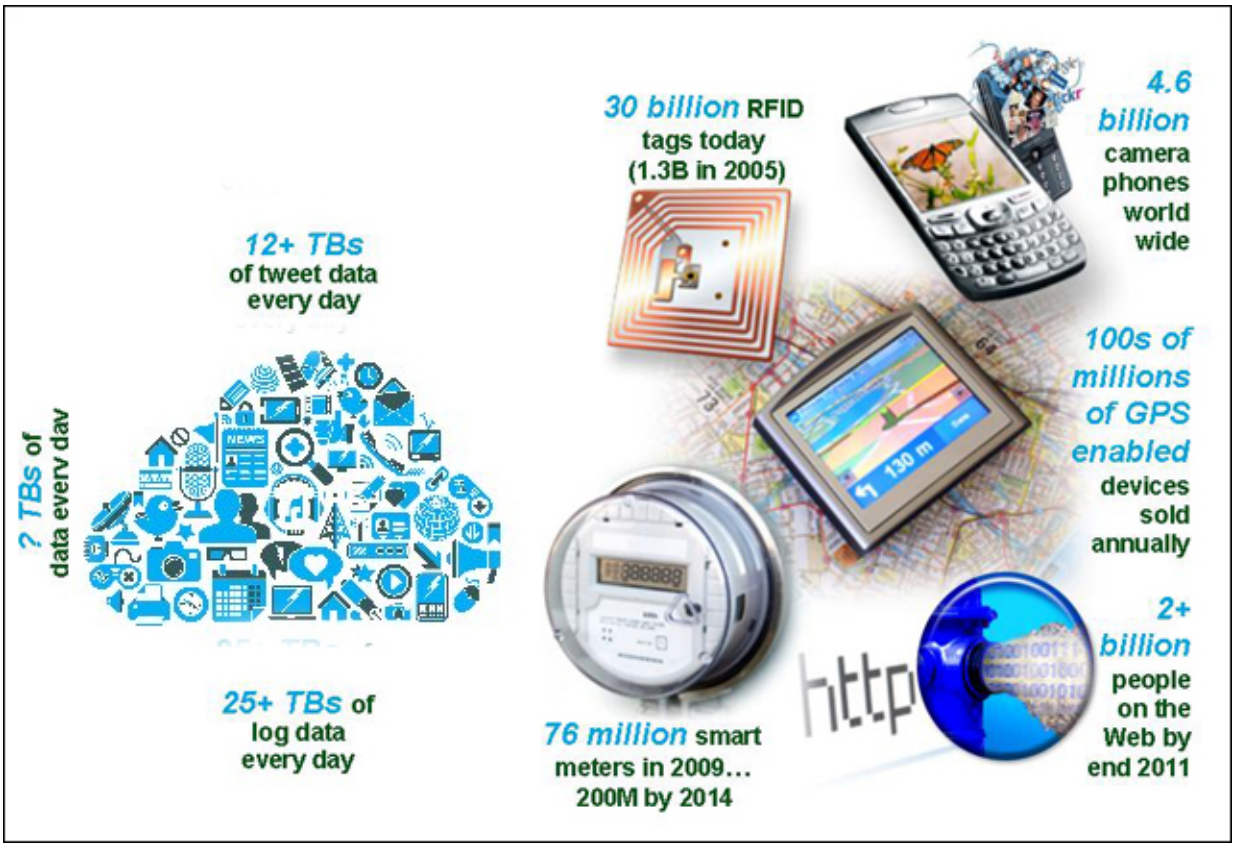

Figure 1. Big data explosion

To provide the capability of using big data sources and analytics of these sources, IBM has developed the IBM InfoSphere® BigInsights™ product. This offering is based on the open source computing framework known as *Apache Hadoop*. This framework provides unique capabilities to the data management ecosystem and further enhances the value of investment in the data warehouse. This IBM Redbooks® Solution Guide describes the value of big data in organizations and how the InfoSphere BigInsights solution helps organizations to handle this data.

Did you know?

With the advent of email, smartphones, social media, sensors, and machine-generated data, significantly more data is generated today than in the past. But big data is not just about the sheer volume of data being created. With a myriad of unstructured sources that create this data, a greater variety of data is now available. Each source produces this data at different rates or what we call *velocity*. In addition, you still need to decipher the veracity of this new information as you would with structured data.

Business value

IBM InfoSphere BigInsights makes it simpler to use Apache Hadoop and to build big data applications. It enhances this open source technology to withstand enterprise demands, by adding administrative, workflow, provisioning, and security features, in addition to best-in-class analytical capabilities from IBM Research. The result is a more developer-friendly and user-friendly solution for complex, large-scale analytics.

By using InfoSphere BigInsights, enterprises of all sizes can cost effectively manage and analyze the massive volume, variety, and velocity of data that consumers and businesses create every day. InfoSphere BigInsights can help increase operational efficiency by augmenting the data warehouse environment. You can use it as an archive that can be queried so that you can store and analyze large volumes of multistructured data without straining the data warehouse. You can also use it as a preprocessing hub so that you can explore your data, determine what is the most valuable, and extract that data cost effectively. In addition, you can use it for ad hoc analysis so that you can perform analysis on all of your data.

The InfoSphere BigInsights offering provides a packaged Apache Hadoop distribution, a simplified installation of Hadoop, and corresponding open source tools for application development, data movement, and cluster management. InfoSphere BigInsights also provides more options for data security, which is frequently a concern for anyone who is contemplating incorporating new technology into their data management ecosystem. InfoSphere BigInsights is a component of the IBM Big Data Platform and, therefore, provides potential integration points with the other components of the platform including the data warehouse, data integration, and governance engines, and third-party data analytics tools. The stack includes tools for built-in analytics of text, natural language processing, and spreadsheet-like data discovery and exploration.

Solution overview

These days, high velocity data sources, such as streaming video or sensor data, continuously send data 24x7. When considering current data warehouse and analytics-intensive environments, data volume is a key factor. Considering that, now and in the future, we will be working with hundreds of terabytes (and in many cases petabytes (PB)), this data has to reside somewhere.

Some might say big data can be addressed by the *data warehouse*. They might suggest that their data warehouse works fine for collection and analysis of structured data and that their solution works well for their unstructured data needs. Although traditional data warehouses do have a role in the big data solution space, they are now a foundational piece of a larger solution. A consideration in data warehouse environments is the I/O that is required for reading massive amounts of data from storage for processing within the data warehouse database server. The ability of servers to process this data is not usually a factor because they typically have significant amounts of RAM and processor power, parallelizing tasks across the computing resources of the server. Many vendors have developed data warehouse appliances and appliance-like platforms (called *data warehouse platforms*), specifically for the analytics-intensive workload of large data warehouses. IBM Netezza® and IBM Smart Analytics Systems are examples of these types of platforms.

Although these data warehouse platforms are optimized for analytics-intensive workloads, they are highly specialized systems and are not cheap. At the rate that data continues to grow, it is feasible to speculate that many organizations will need petabyte-scale data warehouse systems in the next 2 - 5 years. For example, HD video generates about 1 GB of data per minute of video, which translates to 1.5 TB of data that is generated daily per camera. If five cameras are in use, 7.5 TB per day are being generated, which extrapolates to 2.52 PB/year.

You could be adding over 2 PB of data annually to your warehouse that is separate from typical day-to-day, business-centric data systems on which you might already be capturing and performing analytics. The costs of capturing and analyzing big data swell quickly. What if you could use commodity hardware as a foundation for storing data? What if you could use the resources of this hardware to filter data, and then use your existing data warehouse to process the remaining data for its business value? That approach could be more cost effective than expanding your data warehouse platform to a size large enough to perform analytics on all of the data.

Solution architecture

The IBM InfoSphere BigInsights solution is based on the widely used Hadoop framework. Fundamentally, Hadoop consists of two components: a *Hadoop Distributed File System* (HDFS), which provides a way to store data, and *MapReduce*, which is a way of processing data in a distributed manner. These components were developed by the open source community based on documents that were published by Google in an attempt to overcome the problems faced when trying to deal with an overwhelming volume of data. Google published papers on its approach to resolve these issues. Then, Yahoo! started work on an open source equivalent named after a child's toy elephant called *Hadoop*.

Hadoop consists of many connected computers, called *DataNodes*, that store data on their local file system and process the data as directed by a central management node. The management nodes consist of the following processes:

- NameNode

  This process maintains the metadata that relates to where the data is stored on the DataNodes. When a job is submitted, this metadata is accessed to locate the data blocks that are needed by the job. The NameNode is also used, and the metadata is updated if data is saved. No other processing during a MapReduce is carried out on the NameNode. Depending on the version of Hadoop that you are running, the NameNode can be a single point of failure within the Hadoop cluster, and the cluster requires manual intervention if it fails.

- Secondary NameNode

  The Secondary NameNode holds a checkpoint of the metadata on the NameNode and an "edits" file that logs all changes that are made to the locations of the data. This process is not a redundancy for the NameNode but significantly speeds up the process should the NameNode fail.

- JobTracker

  When a MapReduce job is submitted, the JobTracker decides on which nodes the work is to be carried out. The JobTracker coordinates the distributed processing to ensure that nodes local to the data start to carry out the map and reduce functions. Where possible, the JobTracker also ensures that work is carried out simultaneously over multiple nodes.

- TaskTracker

  Each DataNode has a TaskTracker, whose role is to accept jobs from the JobTracker and create a Java virtual machine (JVM) process to do each task.

Figure 2 illustrates the processes within the Hadoop architecture.

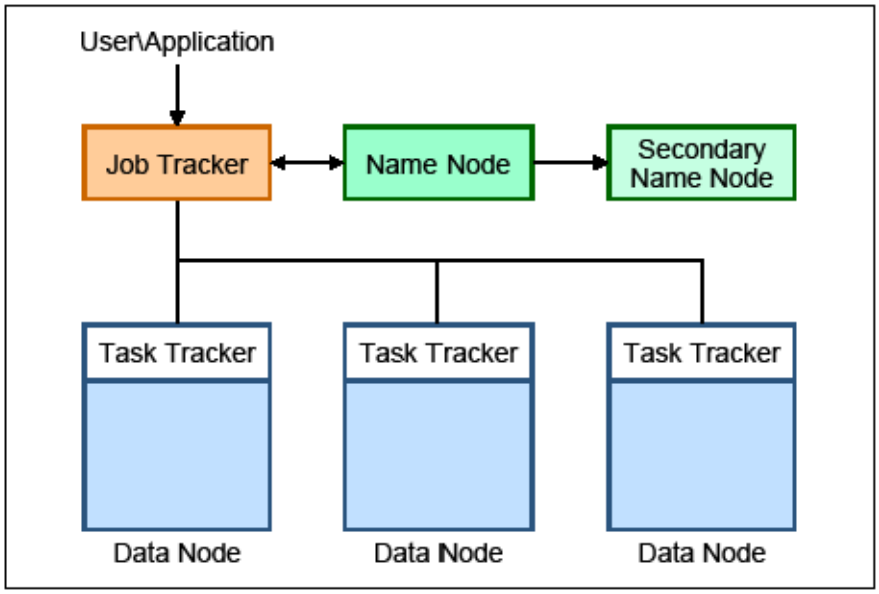

Figure 2. Hadoop architecture

HDFS stores the data in Hadoop by using the following approach:

1. When a file is saved in HDFS, it is broken down into blocks with any remainder data occupying the final block.

   The size of the block depends on the way that HDFS is configured. At the time of writing, the default block size for Hadoop is 64 MB. To improve performance for larger files, InfoSphere BigInsights changes this setting at the time of installation to 128 MB per block.

2. Each block is sent to a different DataNode and written to the hard disk drive (HDD).
3. After the DataNode writes the file to disk, it sends the data to a second DataNode where the file is written.
4. The second DataNode sends the data to a third DataNode.
5. The third node confirms the completion of the write to the second node and then to the first node.
6. The NameNode is then notified, and the block write is complete.

After all blocks are written successfully, the result is a file broken down into blocks with a copy of each block on three DataNodes. The location of all of this data is stored in memory by the NameNode.

Figure 3 illustrates this approach.

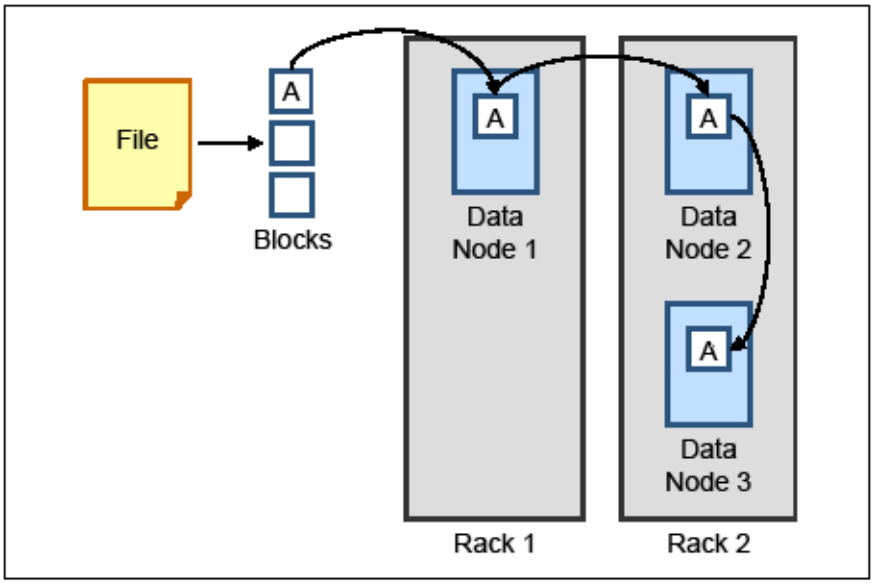

Figure 3. HDFS process

Usage scenarios

If someone says they like your product, what is that statement worth to you? What actions would you take upon learning this information? How much would you pay to learn it? You need a deeper level of understanding, in context, and within a period where responding makes business sense. Today's social media data flow can help.

As an example, consider the IBM big data platform and the results of an effort with a company that makes movies. By using InfoSphere BigInsights as a starting point, a connection is made to multiple social media aggregators. Instead of collecting data manually from each website, the aggregators perform this action from thousands to millions of websites everyday. With this approach, the focus is on the analysis instead of the collection of the data. Comments and details are pulled from the web through these aggregators to collect text-based comments for over 130 movie titles as a baseline.

After the comments are collected, they are processed by the InfoSphere BigInsights text analytics engine to determine sentiment (good or bad) and additional levels of information, including demographics and more detailed topics about the movie trailer. For example, such comments as “the plot was confusing” or “I liked the music a lot” are discovered automatically in the data at rest. The marketing department can use more details, such as “plot” and “music,” to change their message by tweaking the movie trailer before the release of the movie.

When it comes to running an advertisement (movie trailer) during a large event, such as a professional football championship game, the stakes are even higher. Typically, the cost for running an ad is high. If the advertisement (ad) is not received well, the cost can go beyond just the expense of running the ad. If you know in real time that the plot was confusing for ad #1, you might run a tweaked ad #2 later in the show to win your viewers back almost immediately.

Figure 4 provides an overview of the data flow and processing.

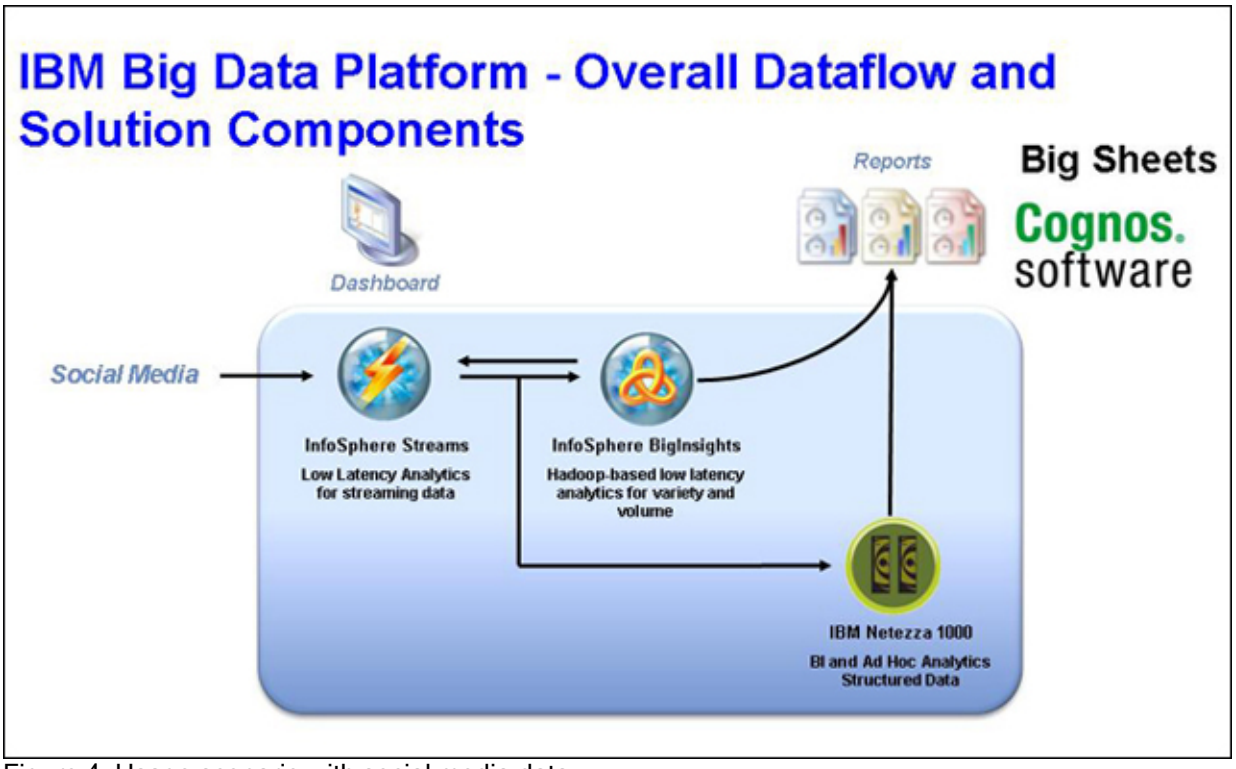

Figure 4. Usage scenario with social media data

The data from the social media aggregators flows into the big data platform. The need to gauge the positive and negative details about the trailer are determined and displayed on a real-time dashboard. Next, the raw data is stored in InfoSphere BigInsights for future processing. This area provides the baseline for the quantity of comments for each detailed topic. Then, depending on the details in the text data, the demographic data is stored in the IBM Netezza 1000 system as structured data. This type of data is analyzed by using IBM Cognos software to produce reports for users to gain a better understanding of the audience that is commenting on their marketing efforts.

This solution is also beneficial in the following use cases:

- Transportation, by incorporating traffic data, weather patterns, logistics, and fuel consumption to optimize travel logistics
- Utilities

  By incorporating weather and energy usage patterns, they can continually update the efficiency of power generation. By monitoring energy usage patterns and weather data, they can decipher the best way to incorporate reusable and renewable energy sources (for example, wind and solar).
- Law enforcement, by implementing for real-time, multimodal surveillance
- Rapid detection of cyber security breaches (computer forensics, real-time monitoring of cameras, situational awareness)
- Information technology (IT)
- Analysis of historical log data to discover past data breaches or vulnerabilities
- Analysis of log data across the entire data center for indications of the overall health of the data center

- Financial services
- Analytics of customer data to determine behavior patterns
- Detection of incidences of potential identity theft or fraud
- Improvement of risk management models through the incorporation of additional data, such as a change in a life situation

Integration

The following products (shown in Figure 4) are among other IBM products that can be used for a comprehensive BigInsights solution:

- IBM InfoSphere Streams

  InfoSphere Streams can perform analytics on continuously streaming data before it lands inside a data warehouse. InfoSphere Streams computing is ideal for high-velocity big data where the ability to recognize and react to events in real time is helpful.

- IBM Cognos

  Cognos can analyze structured data and produce reports for users for a better understanding of events.

- IBM Big Sheets

  IBM Big Sheets is a browser-based analytics tool to extend the scope of your business intelligence data. With Big Sheets, you can easily view and interact with massive amounts of data into consumable, situation-specific business contexts.

- IBM Netezza

  The IBM Netezza 1000 appliance is designed for rapid and deep analysis of data volumes that scale into PB, delivering a performance improvement of 10 - 100 times, at a fraction of the cost of other options.

Supported platforms

This solution is supported by platforms with the following features:

- x86 64-bit systems with a minimum of 4-GB memory
- Minimum 40 GB of disk storage

The following reference architecture is supported:

- Management node: IBM System x3550 M4
- Data node: IBM System x3630 M4

Ordering information

This product is available only through IBM Passport Advantage®. It is not available as a shrink wrapped product.

- License function title: BigInsights
- Product group: IBM InfoSphere

Table 1 shows the ordering information.

Table 1. Ordering part number and feature code

Program name	PID number
IBM InfoSphere BigInsights Enterprise Edition V2.0	5725-C09

Related information

For more information, see the following documents:

- *Implementing IBM InfoSphere BigInsights on System x*, SG24-8077
  http://www.redbooks.ibm.com/abstracts/sg248077.html
- Big data
  http://www.ibm.com/software/data/bigdata
- BigInsights product page
  http://www.ibm.com/software/data/infosphere/biginsights/enterprise.html
- IBM Offering Information page (to search on announcement letters, sales manuals, or both):
  http://www.ibm.com/common/ssi/index.wss?request_locale=en

  On this page, enter `InfoSphere BigInsights`, select the information type, and then click **Search**. On the next page, narrow your search results by geography and language.
- BigInsights Information Center Page
  http://pic.dhe.ibm.com/infocenter/bigins/v2r0/index.jsp

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document was created or updated on January 30, 2013.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips0934.html .

Trademarks