IBM Analytics
Thought Leadership White Paper

June 2015

Using IBM BigInsights for Apache Hadoop to accelerate big data time to value

IBM

Contents

IBM BigInsights for Apache Hadoop overview

IBM® BigInsights™ for Apache Hadoop is a hardware-agnostic software platform that provides new ways of using diverse and large-scale data collections. This white paper describes commonly used capabilities in BigInsights that allow organizations to cost-effectively analyze a wide variety and large volume of data to gain insights that were not previously possible.[1]

BigInsights is focused on providing enterprises with the capabilities they need to meet critical business requirements while maintaining compatibility with the Hadoop project. BigInsights includes a variety of IBM technologies that enhance and extend the value of open-source Hadoop software to facilitate faster time to value, including application accelerators, analytical facilities, development tools, platform improvements and enterprise software integration. While BigInsights offers a wide range of capabilities that extend beyond the Hadoop functionality, IBM has taken an opt-in approach: you can use the IBM extensions to Hadoop based on your needs rather than being forced to use the extensions that come with BigInsights.

To help you initiate big data projects quickly, BigInsights offers a number of enhancements, including a collection of popular open-source and IBM technologies that can be grouped into the following categories:

- Accelerating deployments by tapping into Hadoop community innovation
- Leveraging existing SQL skills and solutions
- Enabling user-driven analytics and data provisioning
- Supporting human-oriented information discovery and topic generation
- Leveraging in-motion and at-rest analytics
- Integrating with popular modeling and predictive analytics solutions

BigInsights: A highly compatible platform

Third-party applications, partner solutions and custom development projects that are compatible with the following BigInsights support versions are expected to work without changes beyond updating data locations.

- Apache Hadoop (1.1.1), a 64-bit Linux version of the IBM SDK for Java 6, and Java
- Avro (1.7.2), a data serialization system
- Chukwa (0.5.0), a data collection system for monitoring large distributed file systems
- Fair Scheduler, for basic management of job submission
- Flume (1.3.0), a distributed, reliable and highly available service for efficiently moving large amounts of data around a cluster
- HBase (0.94.3), a non-relational distributed database written in Java
- HCatalog (0.4.0), a table and storage management service for Hadoop
- Hive (0.9.0), a data warehouse infrastructure that facilitates both data extraction, transformation and loading (ETL), and the analysis of large data sets that are stored in the Hadoop Distributed File System (HDFS)
- IBM BigInsights Jaql, a query language designed for JavaScript Object Notation (JSON), primarily used to analyze large-scale semi-structured data
- Lucene (3.3.0), a high-performance, full-featured text search engine library written entirely in Java
- Oozie (3.2.0), a workflow coordination manager
- Orchestrator, an advanced MapReduce job control system that uses a JSON format to describe job graphs and the relationships between them
- Pig (0.10.0), a platform for analyzing large data sets, consisting of a high-level language for expressing data analysis programs and an infrastructure for evaluating those programs
- Sqoop (1.4.2), a tool that imports information from structured databases and related Hadoop systems into Hadoop clusters
- ZooKeeper (3.4.5), a centralized service for maintaining configuration information that provides distributed synchronization and group services

This paper describes how these enhancements help extend the value of open-source Hadoop with the capabilities organizations need to cost-effectively support emerging big data workloads.

Accelerating deployments by tapping into Hadoop community innovation

The IBM commitment to the Hadoop open-source software components in BigInsights helps facilitate third-party interoperability and supports the ongoing development of new features and functionality. Organizations with existing MapReduce, Hive, Pig and Sqoop projects can leverage that work on BigInsights if the version levels are all compatible and directory structures are mirrored.

Leveraging existing SQL skills and solutions

Legacy applications depend on SQL to access stored data, and SQL is the de facto language used to query structured data; as a result, most organizations have deep and abundant SQL skills. IBM customers have been asking for ways to leverage their SQL skills with Hadoop to lower the barrier to getting started with Hadoop and to facilitate interoperability with existing SQL-oriented tools and applications. IBM is enabling customers to do exactly that with the introduction of IBM Big SQL, a data warehouse system for Hadoop that is used to summarize, query and analyze data that is stored in BigInsights.

Big SQL uses JDBC or ODBC drivers to access data that is stored in BigInsights in the same way that users access databases from their enterprise applications. You can use the Big SQL server to execute standard SQL queries, and to execute multiple queries concurrently (see Figure 1).

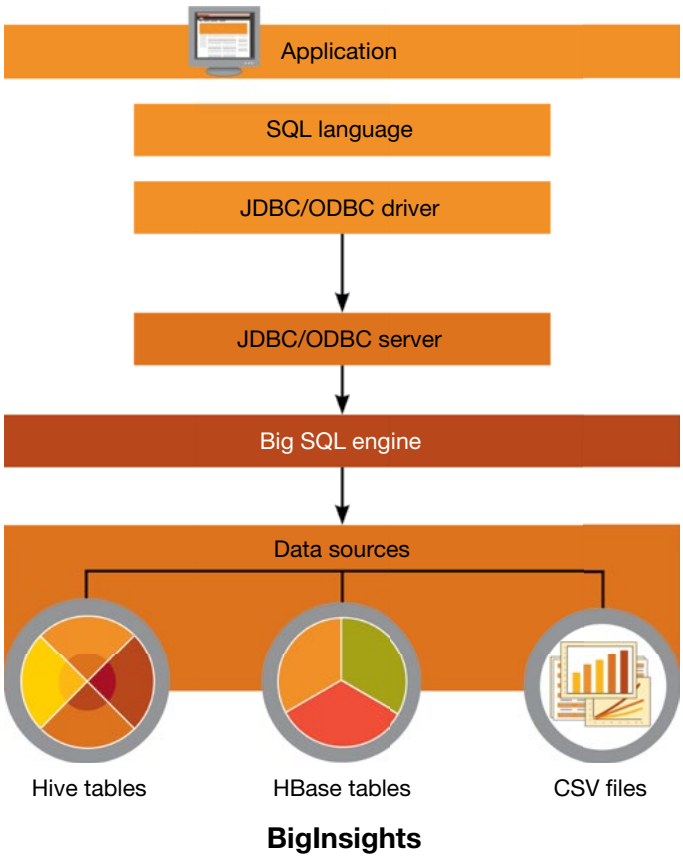

Figure 1. An overview of IBM Big SQL.

Big SQL provides support for large ad hoc queries by using MapReduce parallelism and point queries, which are low-latency queries that return information quickly to reduce response time and provide improved access to data. The Big SQL server is multi-threaded, so scalability is limited only by the performance and number of CPUs on the computer that runs the server. If you want to issue larger queries, you can increase the hardware performance of the server computer that Big SQL runs on, or chain multiple Big SQL servers together to increase throughput.

Big SQL enables anyone with existing SQL skills to be immediately productive, thereby minimizing project timelines and lowering the financial commitment to a given project. With Big SQL, all your data is SQL-accessible, allowing you to choose the storage format best suited for your application.

Enabling user-driven analytics and data provisioning

To gain new insights and improve business results, you need an environment that is well suited to exploring and discovering data relationships and correlations. With the right technology, you can extend the value of your data warehouse by bringing in new types of data and driving new types of analysis. One of the most common deployment patterns for BigInsights is known as a *Data Exploration Zone*. The Data Exploration Zone provides an environment with the capabilities you need to analyze information in its raw form—whether it is structured or unstructured data—with tools such as text analytics, data mining, entity analytics and machine learning. You could use the data in this zone for exploratory analytics, or send it to a data warehouse for deeper analytics, providing greater flexibility in how you work with and analyze data.

To deliver the correct information as rapidly as possible, the data warehouse supporting these systems must be optimized for the right balance of analytics performance and operational query throughput. BigInsights provides multiple user-driven capabilities for creating new data collections from raw data sets, expanding existing data sets from traditional relational sources, and performing ad hoc analytics against the data without requiring assistance from IT.

Data can be cleansed, transformed, integrated and aggregated to prepare it for utilizations. Once prepared, data can be published to common constructs such as Hive, or can be staged for use by an external solution such as IBM PureData™ System for Analytics.

IBM BigSheets

IBM BigSheets is a browser-based analytic tool designed to break large amounts of data into consumable, situation-specific business contexts. Easily accessed from the BigInsights console, BigSheets can collect data from multiple sources, including web crawling, data import/export, data sampling, social media data collection and analysis, machine data processing and analysis, ad hoc queries and more (see Figure 2). BigSheets can also be used with data that is loaded through other means such as Flume or IBM InfoSphere® Information Server.

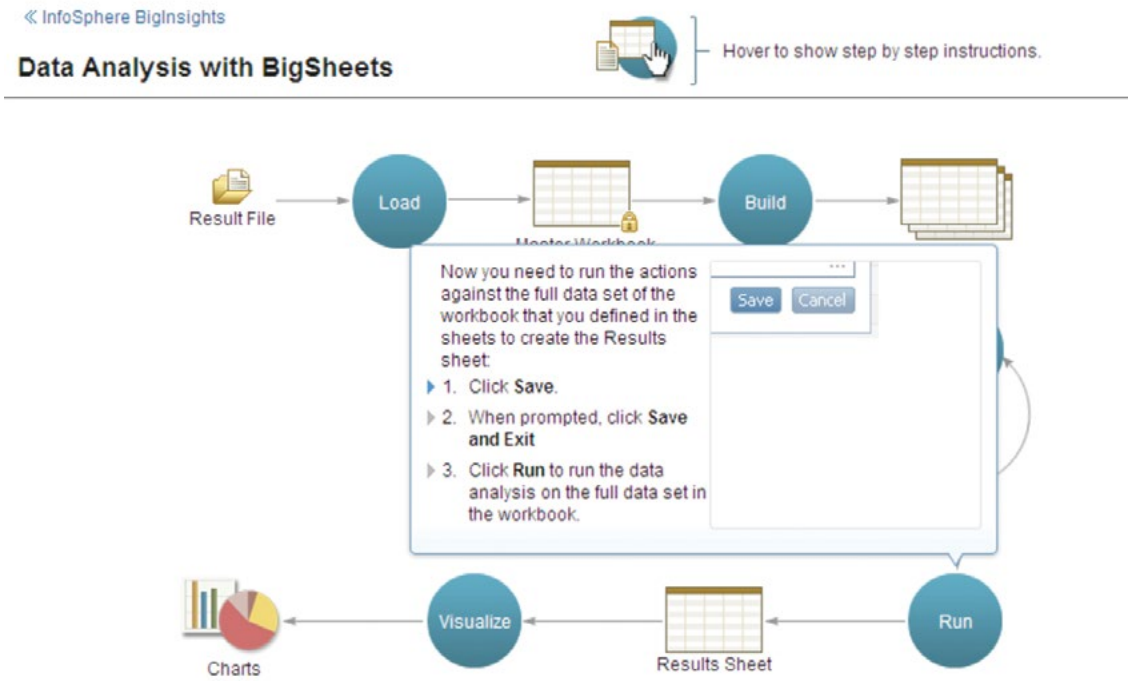

Figure 2. Overview of IBM BigSheets.

Once BigInsights collects the data, BigSheets users load the data of interest into a *master workbook*. From there, BigSheets allows you to format and explore the data by building *sheets* (which resemble spreadsheets) in *workbooks* that are based on the master workbook. You can combine columns from different workbooks, run formulas and filter data. These manipulations form the basis of your analysis.

BigSheets generates and executes the code necessary to perform all of the data manipulation work automatically, allowing you to work in a visual paradigm rather than working at a lower scripting or Java level. You can also combine data with text analytics functions within BigInsights to filter and manipulate data and drill further into information to derive valuable insights out of raw data.

After refining the data and running the analytics, you can apply visualizations such as tag clouds, bar charts, maps and pie charts. These visualizations provide a consumable output for your data that highlights relationships and distills insights from previously disconnected data.

BigInsights Web Console

A role-based Welcome page within BigInsights dynamically populates data collections and jobs for each user. The software includes applications that can be used for completing various data management tasks. These preinstalled applications have the same properties as applications you create and can be used as jumping-off points for big data projects. Users can create and share new jobs as they are developed, making the Web Console an increasingly powerful starting point for working with BigInsights. The following list represents just some of the applications provided out of the box:

- Ad hoc Hive Query: Use the Ad hoc Hive Query application to create your own customized Hive queries to analyze your data.
- Ad hoc Jaql Query: Use the Ad hoc Jaql Query application to create your own customized Jaql queries to analyze your data.
- Ad hoc Pig Query: Use the Ad hoc Pig Query application to create your own customized Pig queries to analyze your data.
- Ad hoc R Script: This application is used to run an R script. Because Oozie assigns the R script to run on a less-used cluster node, R script must be installed on all nodes of your cluster. The R script reads from and writes to files on local rather than HDFS directories. The Ad hoc R Script application, therefore, can copy input files into required local directories and move output files to HDFS directories.
- BoardReader: The BoardReader application searches for, locates and displays information from multiple web sources, such as online forums, message boards, blogs, news sources and videos.
- Data Download: This application is used to download data from the IBM developerWorks® resource for developers. After you agree to the developerWorks terms and conditions, you can then select a sample data set from the data set drop-down list or access some other data set by entering a URL.
- Data Sampling: Given a large data set and parameters, this application generates a representative data sample. The application samples the input data by using a uniform random sample (without replacement). The application outputs the results to a file whose format is the same format as the input file.
- Data Subset: The Data Subset application is used to create a subset of your complete data. You can then improve performance by analyzing the data subset by structure, content and file format.
- Database Export: This application writes data from files on the HDFS to a table in the relational database management system, and uses a Java program to export the data that is stored on the HDFS into a table in the database. The input data is stored in files on DFS. The format of the input can be CSV or JSON.

- Database Import: This application loads data from a relational database management system into a file on the HDFS. It uses a Java program to import data from a database and write it to a file on the HDFS. You can specify the SQL (Select) query to define the data that is to be imported from the database. The data that is retrieved from the database is then written out to a file on the HDFS in CSV or JSON format.
- Distributed Copy: By using a MapReduce job, you can copy data from a remote source to the HDFS or from the HDFS to a remote source. Use the Distributed Copy application to copy files and directories from one source to another.
- HBase: The HBase application enables you to export rows of data from an HBase table through the BigInsights console. You can export rows of data from an HBase table as a JSON file. The application requires parameters to export the data; it does not accept HBase queries from you.
- Web Crawler: The Web Crawler application is an automated program that methodically tracks Internet pages and collects data. It also compares the size and contents of a file against the version of that file stored in BigInsights.
- Web REST Import: This application fetches content from a specified URL location and stores the content in a specified HDFS directory.

These applications can be modified and then published to specific users or a class of users based on their security rights to provide them with multiple options for starting their projects.

Industry-standard Hadoop with enterprise features

IBM is a founding member of the Open Data Platform Initiative (ODP), an industry association to help drive collaboration, innovation and standardization across Hadoop and big data technologies.

ODP aims to accelerate the delivery of big data solutions by providing a well-defined core platform to target. The consortium will identify a common core including HDFS, and test, certify and standardize the core components of a new open data platform of select Apache Software Foundation (ASF) projects to provide a foundation that big data solution providers can build upon.

Analytics accelerators

IBM offers several analytic accelerators that greatly reduce the time to value of your big data applications. These accelerators provide business logic, data processing capabilities and visualization for specific use cases. By using accelerators, you can apply advanced analytics to help integrate and manage the variety, velocity and volume of data that constantly enters your organization. Accelerators also create a development environment for building new custom analytic applications that are tailored to the specific needs of your organization.

Two accelerators are shipped with BigInsights: IBM Accelerator for Machine Data Analytics and IBM Accelerator for Social Data Analytics. Two accelerators are shipped with InfoSphere Streams: IBM Accelerator for Social Data Analytics and IBM Accelerator for Telecommunications Event Data Analytics. These accelerators cover many common use cases and are easily extended for enterprise-specific needs.

IBM Accelerator for Social Data Analytics

Data from social media forums contains valuable information about user preferences. However, accessing and working with that information requires large-scale import, configuration and analysis capabilities. The IBM Accelerator for Social Data Analytics, which has a built-in understanding of how to work with social data sources, extracts relevant information from tweets, boards and blogs and then builds social profiles of users based on specific use cases and industries. A typical workflow consists of importing your data files and then configuring, indexing and analyzing the data.

With IBM Accelerator for Social Data Analytics, you can:

- Import and analyze social media data, identifying user characteristics such as gender, locations, names and hobbies
- Develop comprehensive user profiles across messages and sources
- Associate profiles with expressions of sentiment, buzz, intent and ownership around brands, products and companies

The IBM Accelerator for Social Data Analytics is commonly used to collect data to enrich customer analytics—and unlike many social listening tools, can be used to help identify social activity down to a known customer.

IBM Accelerator for Machine Data Analytics

The IBM Accelerator for Machine Data Analytics can ingest, parse and extract a variety of machine data from sources such as machine data files, log files, smart devices and telemetry, and help process that data in minutes instead of days or weeks. It helps organizations gain insights into operations, customer experiences, transactions and behavior that may identify infrastructure issues and changes in customer preferences, or trap events that can drive systems of engagement. Many IBM customers use the IBM Accelerator for Machine Data Analytics to proactively boost operational efficiency, troubleshoot problems, investigate security incidents and monitor end-to-end infrastructure to avoid service degradation or outages.

A typical introductory workflow consists of organizing and importing batches of data, and then extracting, indexing, searching, transforming and analyzing the data. With IBM Accelerator for Machine Data Analytics, you can:

- Search within and across multiple machine data entries based on a text search, faceted search or a timeline-based search to find events
- Enrich the context of machine data by adding and extracting log types into the existing repository
- Link and correlate events across systems
- Uncover patterns

BigInsights Text Analytics

BigInsights Text Analytics is a powerful and declarative information extraction system that excels at creating structured information from text inputs, allowing users to gain actionable insights from the underlying text data. The BigInsights Text Analytics module was custom-designed to leverage a Hadoop-oriented processing model. It is extremely fast and capable of handling large amounts of unstructured information very quickly compared to traditional text analytics approaches. The BigInsights Text Analytics module is also declarative, meaning it can be readily adapted to your specific analytics needs using a SQL-like method that is simply not possible with conventional text tools. This helps lower costs as well as providing a level of comfort that is unique in the Apache Hadoop space.

Text Analytics is included in the Eclipse development environment as part of the BigInsights Text Analytics Workflow perspective. Text Analytics Eclipse development tools can be used to develop and test extractors in Eclipse.

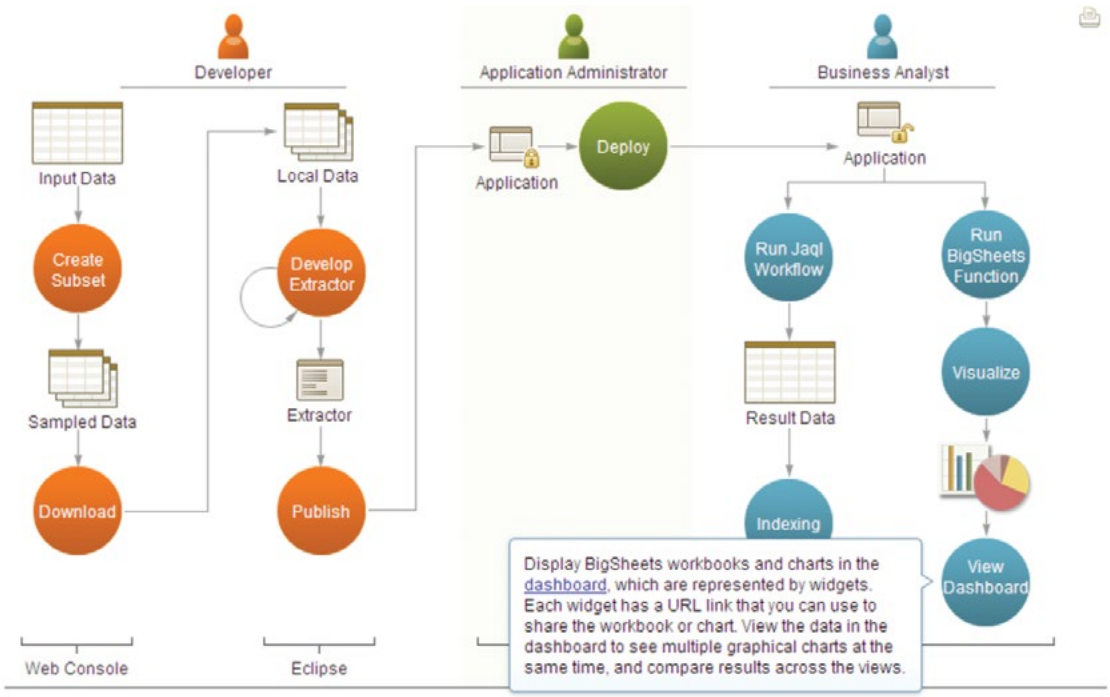

Figure 3. Overview of the BigInsights Text Analytics Workflow.

Once you have identified and selected the extractor you would like to use, you can publish it into the BigInsights Console as an application, which an administrator can then deploy and make available for consumption by users across BigInsights.

The Welcome page of the BigInsights Console includes information about how to enable your Eclipse environment for developing applications using BigInsights. Once the extractor is published and the application is deployed in the BigInsights console, it can then be run as a BigSheets function or as part of a workflow.

Results from a Text Analytics application can be exported to BigSheets, Dashboard and other BigInsights components for further analysis (see Figure 3).

IBM InfoSphere Data Explorer

A key part of your data exploration activities is allowing for human-driven exploration and rapid understanding of the information you have on hand. It facilitates end-user-driven creation of topics of interest and automatic discovery of related information, and lets users quickly build and deploy interactive web applications that are commonly used in customer insight and customer service environments. InfoSphere Data Explorer Engine servers can receive data in

real time from a cluster of BigInsights servers or InfoSphere Streams servers. InfoSphere Data Explorer can also push relevant data to users of information applications and also enables federated access to other IBM products.

Leveraging in-motion and at-rest analytics

Organizations are increasingly deploying analytics and applications that span in-motion (or real-time) and at-rest use cases. The generation of insight that spans in-motion and at-rest use cases requires data and analytics to flow across both types of environments. BigInsights accelerates the ability to span in-motion and at-rest information handling by leveraging IBM InfoSphere Streams.

InfoSphere Streams is a high-performance computing platform that allows user-developed applications to rapidly ingest, analyze and correlate information as it arrives from thousands of real-time sources. For streaming data, InfoSphere Streams can continuously analyze massive amounts of data with very low latency, enabling you to quickly react to trends and events as they unfold. Programmers can instruct InfoSphere Streams to write data as needed to BigInsights for deep analysis of trends over time. The results of this analysis can be captured and fed back to InfoSphere Streams to fine-tune application logic and actions. To reduce deployment time and costs, InfoSphere Streams applications map naturally to distributed BigInsights stores.

Integration with popular modeling and predictive analytics solutions

IBM customers have been asking for easy ways to use a variety of predictive modeling solutions with Hadoop-based discovery zones to quicken time-to-value and better utilize existing solutions and skills. BigInsights meets this request by supporting the most common modeling and analytics packages available, including SAS, IBM SPSS® and R. Users familiar with these modeling environments can use data in BigInsights for their analysis. This allows users to continue to work in a familiar environment while gaining access to new, richer information than previously possible.

Each analytics package has different levels of Hadoop support within the BigInsights environment. SAS and other environments mainly use BigInsights as a data-sourcing environment, allowing you to utilize information in structures such as Hive. Some packages, such as SPSS Catalyst for BigInsights, allow for model development and execution to be done directly on the BigInsights platform.

SPSS Catalyst enhances analytical productivity and shortens time to value by helping to automate portions of data preparation, automatically interpreting results and presenting analyses in interactive visuals with clear, concise summaries. SPSS Analytic Catalyst running with BigInsights allows automated key driver identification with sophisticated algorithms, automatic testing and regression-based techniques. SPSS Analytic Catalyst also provides interactive visuals and plain language summaries of predictive analytics findings that show insights at a glance, along with supporting explanations and statistical details.

Conclusion

BigInsights provides a unique set of capabilities that combine the innovation from the Apache Hadoop ecosystem with robust support for traditional skill sets and already-installed tools. The ability to leverage existing skills and tools through open-source capabilities helps drive lower total cost of ownership and faster time to value.

For more information

To learn more about BigInsights and the BigInsights for Apache Hadoop Quick Start Edition, please contact your IBM representative or IBM Business Partner, or visit the following websites:

- **ibm.com**/software/data/infosphere/biginsights
- **ibm.com**/infosphere/quickstart

About the author

Tom Deutsch (@thomasdeutsch) serves as CTO for the IBM Industry Solutions team. He played a formative role in the transition of Hadoop-based technology from IBM Research to the IBM Software Group, and he continues to be involved with IBM Research big data activities as well as transitions from IBM Research to commercial products. Deutsch created the BigInsights for Apache Hadoop product and spent several years helping customers with Hadoop, BigInsights and InfoSphere Streams technologies, including identifying architecture fit, developing business strategies and managing early-stage projects across more than 200 customer engagements. With more than 20 years in the industry and as a veteran of two startups, Deutsch is an expert on the technical, strategic and business information management issues facing the enterprise today.

IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
June 2015

[1] This paper does not cover all of the more than 30 capabilities that BigInsights contributes beyond what is available with open-source Hadoop distributions. For comprehensive product documentation on all BigInsights features, please see the Product Center documentation at http://www-01.ibm.com/support/knowledgecenter/SSPT3X/SSPT3X_welcome.html

Please Recycle

IMW14684-USEN-03