Exploring your InfoSphere BigInsights cluster and sample applications

Quick start with the web console

Cynthia M. Saracco (saracco@us.ibm.com)
Senior Software Engineer
IBM

Priya Baliga
Advisory Software Engineer
IBM

Stephen A. Brodsky
Distinguished Engineer, Architect, Big Data
IBM

Skill Level: Introductory

Date: 12 Apr 2012

If you're looking to getting a quick start with "big data" projects involving IBM® InfoSphere® BigInsights, you'll want to become familiar with its integrated web console. Through this tool, you can explore the health of your cluster, navigate your distributed file system, launch IBM-supplied sample applications, monitor the status of jobs and workflows, and analyze data using a spreadsheet-style tool. This article takes you on a tour of the Web console, highlighting key capabilities that can help you get up to speed quickly.

About InfoSphere BigInsights

InfoSphere BigInsights 1.3 is a software platform designed to help companies discover and analyze business insights hidden in large volumes of a diverse range of data — data 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 businesses derive value from such data in an efficient manner, the Enterprise Edition of BigInsights includes several open source projects (including Apache Hadoop) and a number of IBM-developed technologies. Hadoop and

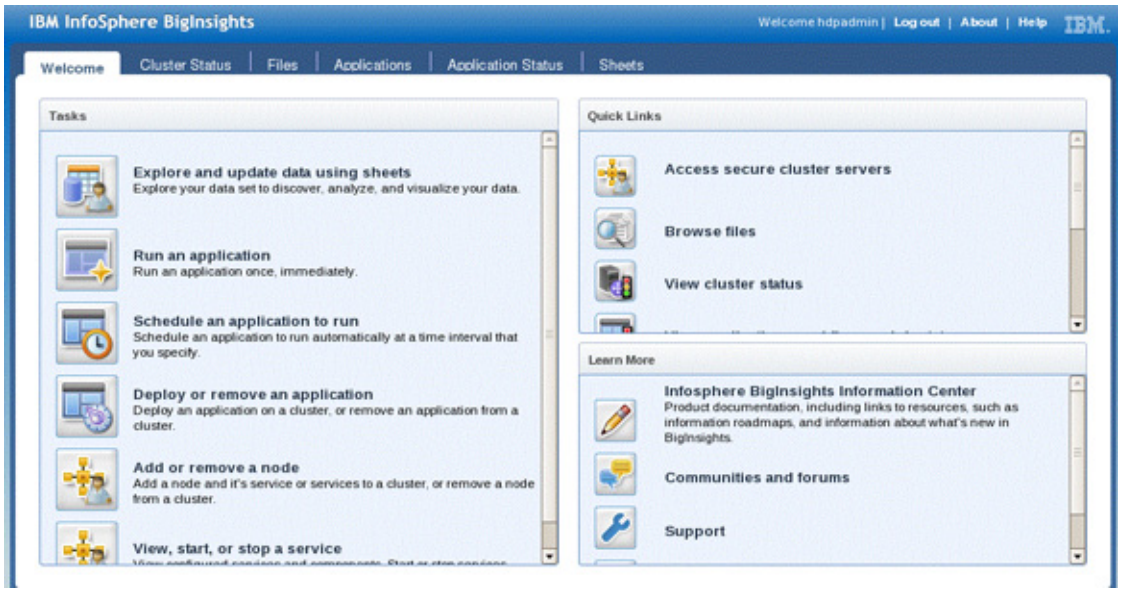
its complementary projects provide an effective software framework for data-intensive applications that exploit distributed computing environments to achieve high scalability. IBM technologies enrich this open source framework with analytical software, enterprise software integration, platform extensions, and tools. For more on BigInsights, see Resources. This article focuses on one IBM-specific technology included with BigInsights 1.3 Enterprise Edition: its web console.

As you'll see, the web console includes tools for administrators, application developers, and business analysts. In addition, the web console can also help you secure your cluster by limiting the number of open ports and supporting LDAP or file-based authentication.

First steps

Once BigInsights is running, you can easily launch the console from a browser. Simply specify the host name and port number identified at installation time for the web console. For SSL installations, the default is https://<host name>:8443. For non-SSL installations, the default is http://<host name>:8080. After providing a valid user ID and password, the Welcome page of the web console will appear, as shown in Figure 1.

Figure 1. Welcome page for BigInsights 1.3 Enterprise Edition web console



The Welcome page features links for common tasks, such as running applications, adding and removing nodes, and exploring data using a spreadsheet-like tool. In addition, it includes links to popular external resources, such as the BigInsights InfoCenter (product documentation) and community forum.

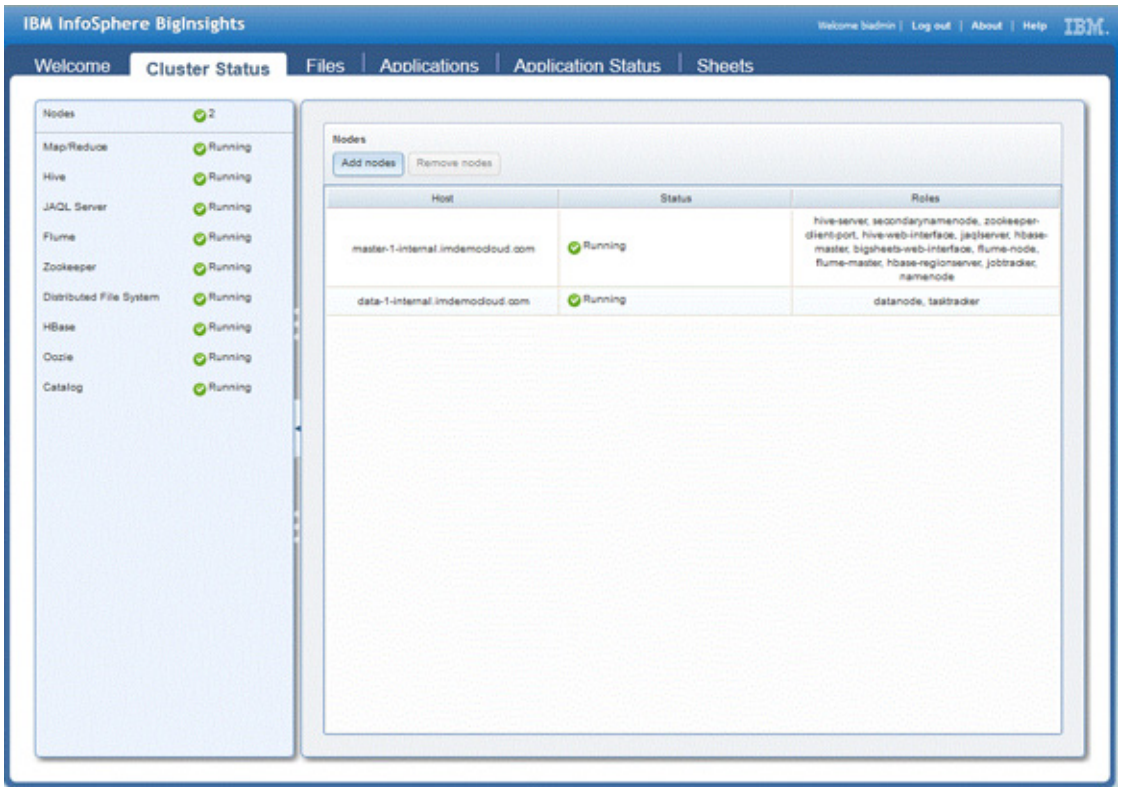
Subsequent sections of this article explore the key capabilities of the console in greater detail. Administrators may be particularly interested in operations available through the Cluster Status, Files, Applications, and Applications Status pages.

Application developers are likely to work most frequently with the Files, Applications, and Applications Status pages. Business analysts may be most inclined to analyze data through BigSheets (a spreadsheet-style tool) or launch published applications through the Applications page. However, they may sometimes want to explore the Files or Applications Status pages as well.

Administering your cluster

Through various web console links, administrators can inspect the overall health of their cluster, as well as perform many common functions, such as starting and stopping specific services, adding nodes, etc. The Welcome and the Cluster Status pages serve as the starting points for most popular administrative operations. For example, the Cluster Status page provides a real-time view of the cluster's health and enables administrators to add nodes to their clusters as needed. Figure 2 depicts the status of a two-node BigInsights test environment in which all services are actively running.

Figure 2. Inspecting the status of a BigInsights environment

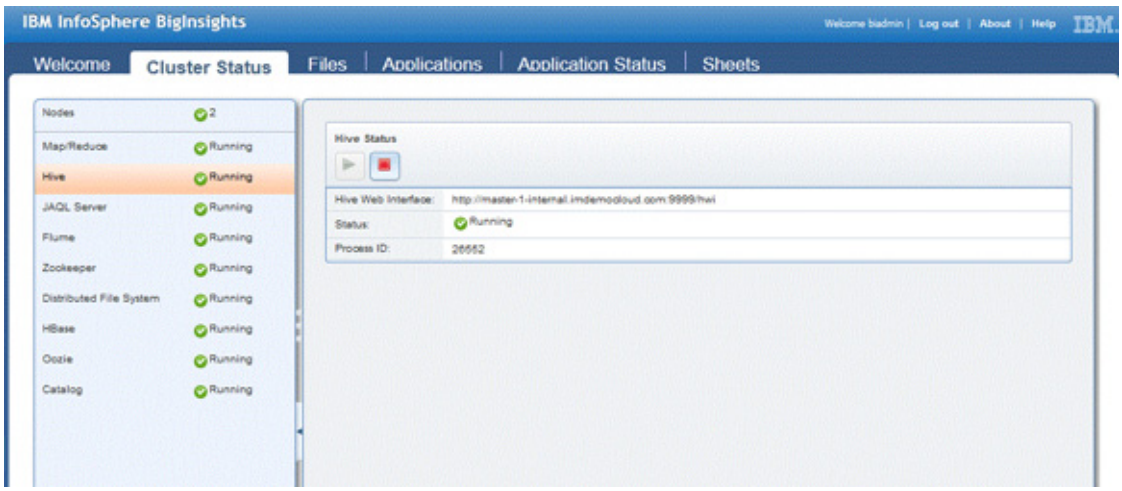


To drill down into the status of any service, administrators simply click on the service of interest in the left pane. The right pane displays detailed information, including the process ID and additional data that varies for each service. In addition, administrators can also use the right pane to start or stop the service identified.

Figure 3 depicts an actively running Hive service, which an administrator can stop simply by clicking on the provided button. In addition, because Hive is an open

source project that includes a web-based interface, the BigInsights console includes the URL for launching it.

Figure 3. Inspecting the status of a specific BigInsights service



And administrators can launch various open source tools through the Welcome page. The **Access secure cluster servers** item in the Quick Links pane provides an easy way to launch tools provided with open source projects, such as Hadoop, Flume, and Hbase. Figure 4 shows the list of displayed links.

Figure 4. Quick links for administering various open source components

URL	Alias
http://master-1-internal.imdemocloud.com:60030	hbase-regionserver
http://master-1-internal.imdemocloud.com:60010	hbase-master
http://master-1-internal.imdemocloud.com:50090	secondarynamenode
http://master-1-internal.imdemocloud.com:50070	namenode
http://master-1-internal.imdemocloud.com:50030	jobtracker
http://master-1-internal.imdemocloud.com:35871	flume-master
http://master-1-internal.imdemocloud.com:35862	flume-node
http://data-1-internal.imdemocloud.com:50075	datanode
http://data-1-internal.imdemocloud.com:50060	tasktracker

Working with your distributed file system

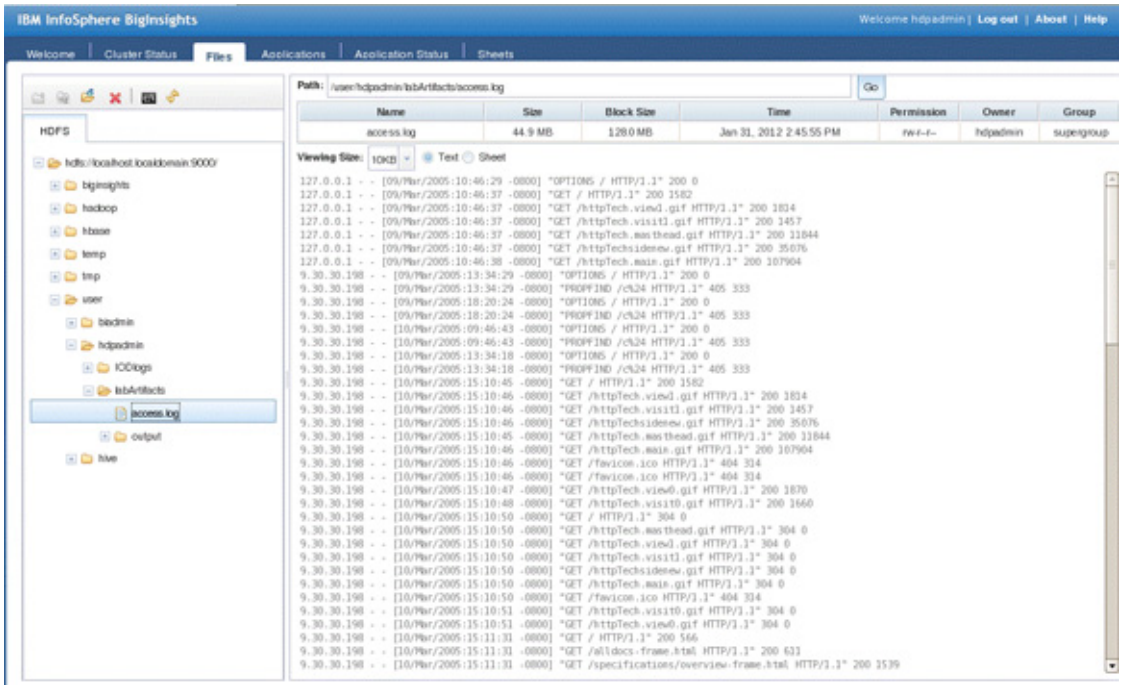
You can also use mechanisms for exploring the Hadoop distributed file system (HDFS) and performing basic file system functions, such as uploading or downloading files, creating and deleting subdirectories, and issuing HDFS shell

commands. Some file system functions are particularly useful for administrators, while others can help users get started with specific analytical projects.

As shown in Figure 5, the Files page of the web console includes a file system navigation tool in the left pane. Icons at the top enable you to create a directory, upload a file to HDFS, download a file from HDFS to your local file system, delete a file or directory from HDFS, open a command window to launch HDFS shell commands, and refresh the web console page. The file upload/download buttons are best suited for working with small test files. To move high volumes of data, consider using HDFS shell commands, the Distributed File Copy sample application (which we'll discuss shortly), or an open source tool like Flume.

The right pane of the Files page displays information about the particular file or directory you've highlighted in the navigation pane. For example, if you navigate to an individual file, the top portion of the right pane displays the file's path, permissions, owner, size, and other details. In the bottom portion, the right pane displays a small subset of the file's contents in text format. In Figure 5, we see the first 10 KB of the access.log file, which contains web log records that can be easily viewed as text.

Figure 5. Working with your distributed file system



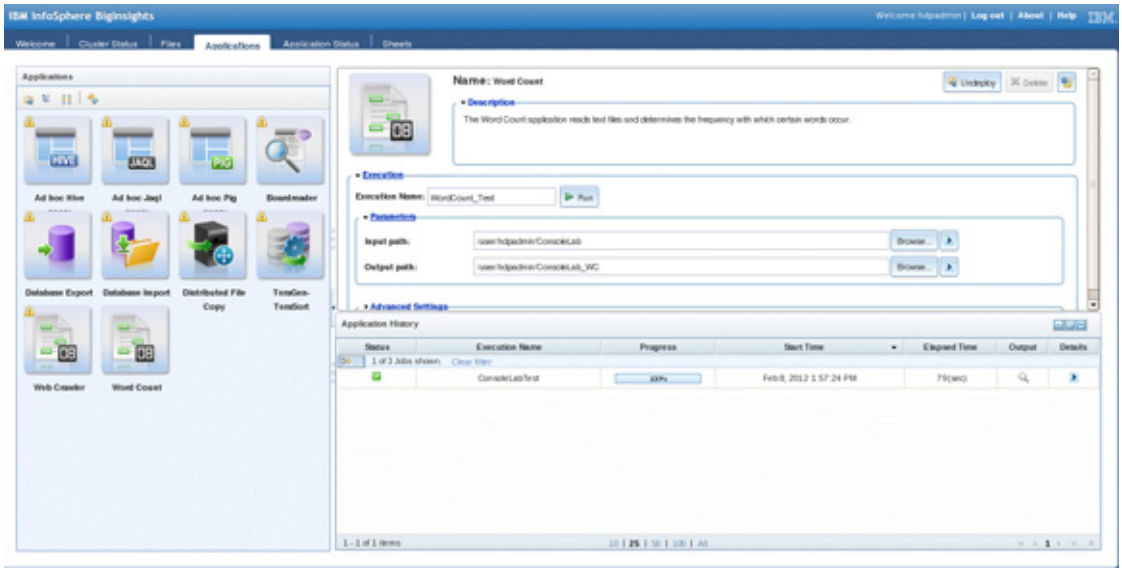
Some types of files can be easily displayed as a "Sheet" (a spreadsheet-style format). You'll see how to do that a little later.

Exploring the application catalog and launching applications

The Applications and Applications Status pages of the console enable you to launch deployed applications, including sample applications provided by IBM; inspect the status of applications and workflows; and review execution details.

As shown in Figure 6, the Applications page enables users and administrators to work with applications that have been uploaded to (i.e., published in) the BigInsights application catalog. The left pane depicts these applications, which include IBM-provided sample query applications, data import/export applications, and test applications. We'll discuss each of these briefly. However, it's worth noting that the upper left corner of each application icon indicates the application's state of readiness. A yellow triangle in the upper left indicates that the application isn't ready for use because it hasn't been deployed on the cluster. An icon without this marker has been deployed and is ready for use. When you first install BigInsights, all sample applications will have a yellow triangle in upper left corner because none will have been deployed. However, deploying these applications — or any custom-written application you upload to the catalog — is a simple matter for application administrators, as you'll see. In Figure 6, only the WordCount sample application has been deployed.

Figure 6. Exploring and launching applications



Sample query applications provided with BigInsights enable developers to dynamically issue Hive, Pig, or Jaql queries. Using the web console can be convenient for prototyping and exploratory work, enabling application developers to quickly test queries and inspect results with minimal effort. By contrast, Eclipse-based plug-ins provided for BigInsights are more appropriate for production-level application development work.

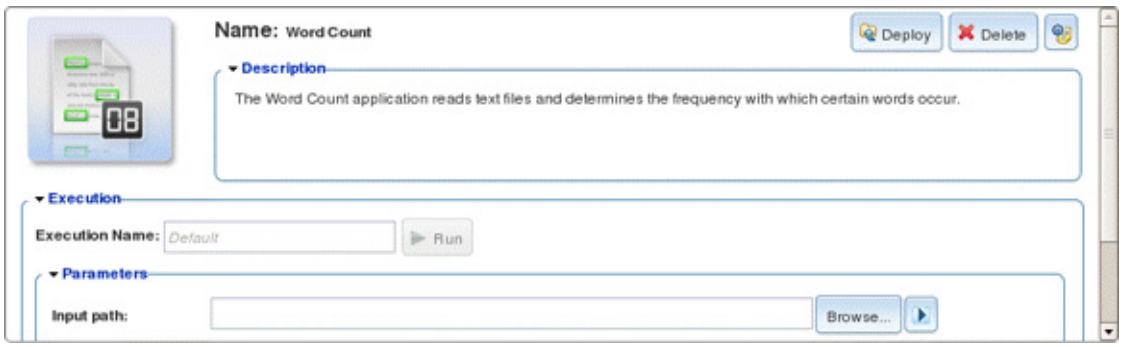
Sample data import and export applications provided with BigInsights support:

- Moving data between a relational database management system (DBMS) and HDFS. Specifically, the Database Import and Database Export applications use BigInsights' Jaql JDBC module to extract data from HDFS into a relational DBMS and vice-versa. Supported DBMS platforms include DB2®, Oracle, Teradata, Informix®, SQL Server, and Netezza.
- Moving data between a remote file system and HDFS using the Distributed File Copy sample application.
- Conducting web searches and obtaining qualifying web data. The Web Crawler sample application uses open source Nutch technology to search the web.
- Conducting searches of public forums, videos, micro-blogging sites, and other web-based communities. The Boardreader sample application uses the search APIs supported by Boardreader.com to obtain qualifying results spanning various websites. (Users must obtain valid software license keys from Boardreader.com to execute this application.)

Finally, BigInsights includes two sample test applications popular in Hadoop-based environments: WordCount and TeraGen-TeraSort. WordCount processes a collection of text files, returning the total of the number of occurrences of each word found. TeraGen-TeraSort generates and sorts terabyte-sized data sets.

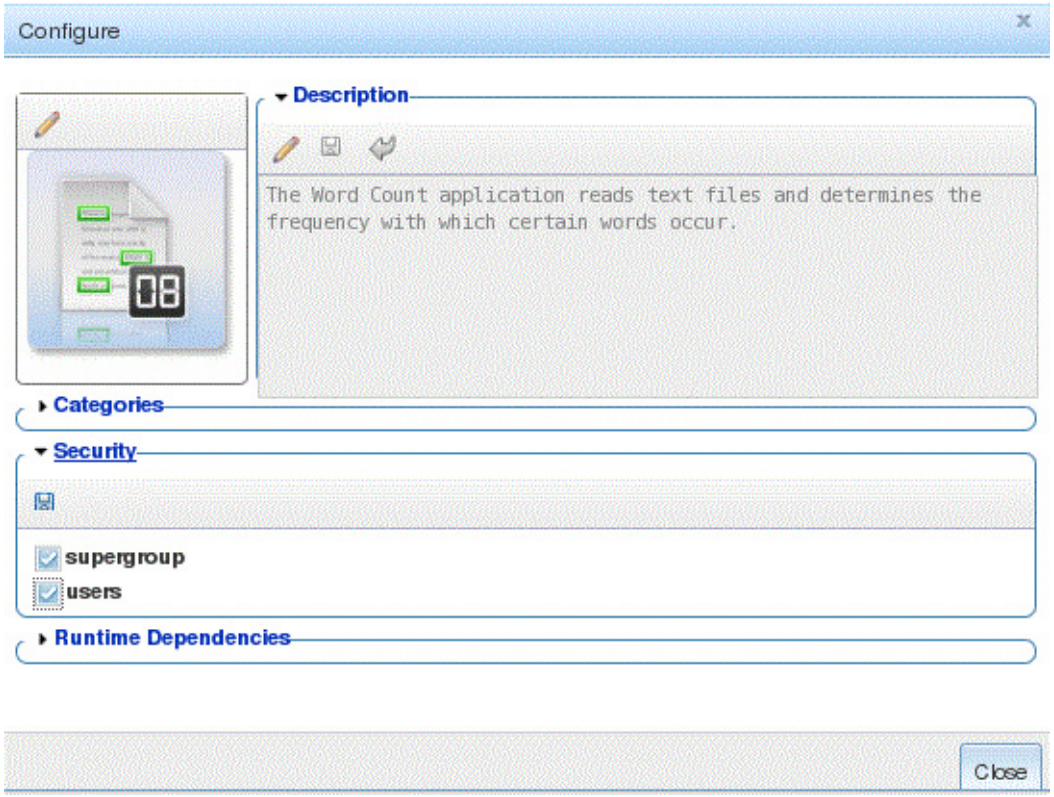
As mentioned, before a sample application (or user-written application) that's been published in the catalog can be used, it must be deployed to the BigInsights cluster. To do so, an administrator clicks on the application's icon, and the right pane displays options for deploying, deleting, and configuring the application, as shown in Figure 7.

Figure 7. Deploying an application



The **Configure** button (shown in the upper right corner of Figure 7, next to the **Delete** button) allows administrators to specify who's authorized to launch the application. For example, the settings shown in Figure 8 indicate that members of the "supergroup" and "users" groups will be authorized to access the application once deployed. (After logging into the console, end users will only see applications they're authorized to launch.) After configuring the application, the administrator simply clicks on the **Deploy** button to make the application available to authorized users.

Figure 8. Configuring security characteristics of an application published in the web console's catalog



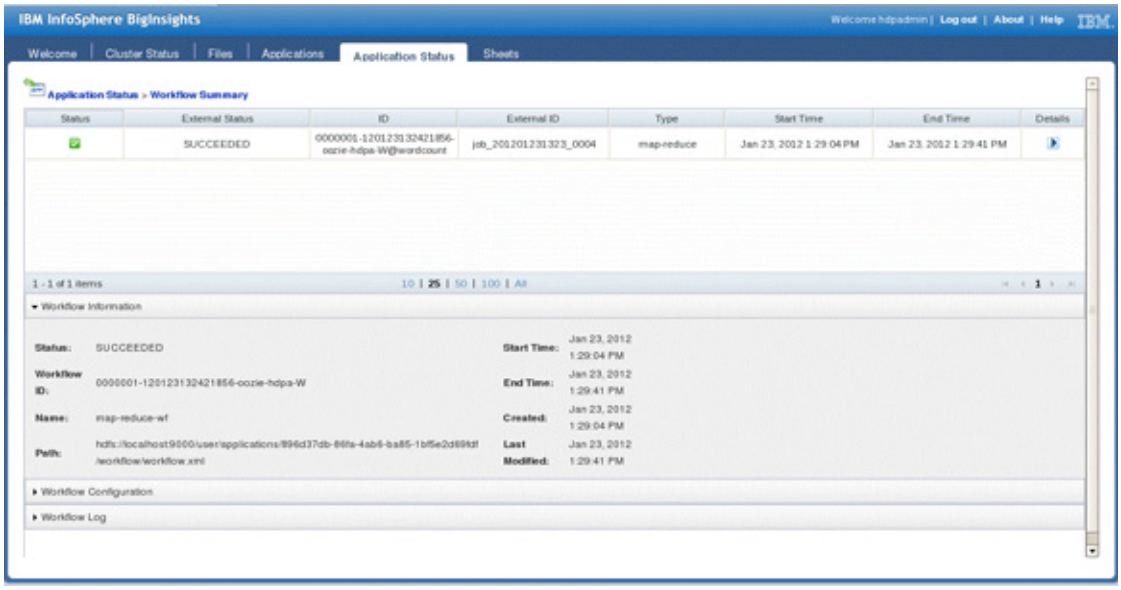
Applications frequently require input and output parameters, which users can specify in the right pane at launch time (see Figure 6). After providing required parameters and an execution name for the application, a user can run the application and monitor its status in real time by reviewing information displayed in the Applications History pane at lower right. While the application is running, a **Stop** button will become active, allowing users to terminate the application if desired.

As mentioned, programmers can publish their own applications to the catalog for subsequent deployment on the cluster. Graphical wizards provided with the BigInsights Eclipse plug-ins guide programmers through the process of identifying their target application, specifying a workflow configuration file (or accepting a generated file), providing details about the application's parameters, and creating a ZIP file that will be uploaded to the target BigInsights server.

Monitoring workflow and application status

The BigInsights web console generates an Oozie-based workflow for each application, and users can inspect details about the workflow and its associated jobs. For example, Figure 9 depicts details about a successfully executed workflow, including its start and end time, its ID, and other data.

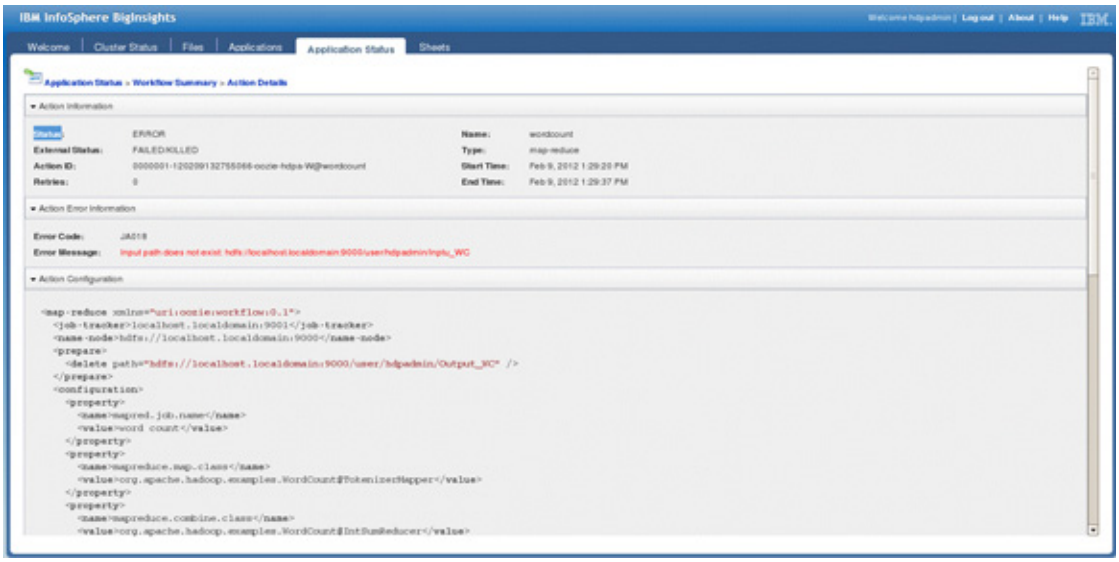
Figure 9. Inspecting the status of a completed application workflow



Further details about the job are available through provided links. For example, you can determine the number of setup, map, and reduce tasks required for your job; review configuration data; examine statistical data about your job (such as number of bytes read and written); and inspect log data.

Exploring details about your workflow or job can often help you diagnose runtime errors. Figure 10 displays the Action Details associated with an application that failed to run successfully. A quick examination of the data indicates that the application — in this case, a WordCount run — could not locate the specified input directory of hdfs://localhost.localdomain:9000/user/hdpadmin/Inptu_WC. (Most likely, the invoker meant to reference `../Input_WC` as the input directory.) With this information, it's a simple matter to correct the input directory and re-run the application.

Figure 10. Examining diagnostic information returned from a failed application run



Using a spreadsheet-style tool to analyze and explore your data

The Sheets page enables users to explore and analyze big data using a spreadsheet-style interface called BigSheets. Unlike like many other big data tools, it's designed for business analysts and non-technical professionals. With BigSheets, business users model data stored in the BigInsights distributed file system as *sheets* or *collections*.

Typically, users filter, explore, and enrich the contents of their collections using built-in functions and macros. Furthermore, some users combine data residing in different collections, creating new sheets (collections) and charts to visualize their data. Finally, users can export the results of their BigSheets analyses into a variety of common formats for use by downstream applications. IBM provides export facilities for HTML, JSON, CSV, RSS, and ATOM data.

A full discussion of BigSheets is beyond the scope of this article, so we'll briefly walk through a sample scenario that illustrates one way in which this tool can be used. Companies can collect data from websites, files and other sources into BigInsights using a variety of tools and techniques. Examples of data collection and import mechanisms include Flume, HDFS shell commands, and sample applications accessible through the web console. Users can explore and manipulate the data using BigSheets, also accessible through the web console.

To create a collection, users can work through the Files page to identify the data of interest, specifying a viewing preference of "Sheet." After doing so, the console will prompt the user to specify an appropriate "reader" or data format translator. IBM provides several built-in readers for working with common data formats, including CSV, TSV, web crawler data, JSON, and others. In addition, Java™ programmers

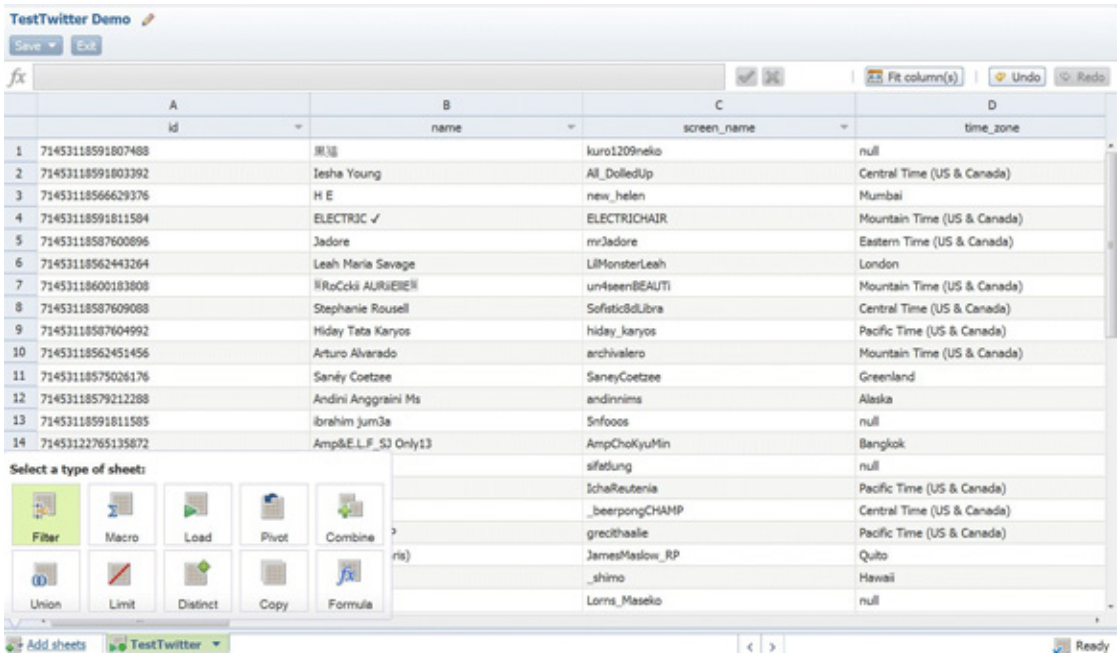
can create custom plug-ins to handle specific data formats and make these available to business users of BigSheets.

After saving the sheet, users can employ built-in functions and macros to customize their collections. For example, basic editing functions include renaming columns, inserting new columns, deleting columns, and sorting data. More sophisticated data manipulation functions include using built-in operators to filter data, define formulas, apply macros, combine data from multiple collections, etc. In addition, Java programmers can create plug-ins that provide additional functions and macros if needed.

As the user tailors the content of the collection through the Sheets graphical editor, BigInsights translates these commands into executable scripts run against a subset of the data represented by the collection. This supports exploratory, iterative analysis in a timely fashion. Once the user is satisfied with the changes made to the collection, he clicks a button to instruct BigInsights to run the collection against the full set of data it represents. Depending on the data volumes involved, this may take some time, so BigSheets provides a real-time status bar that indicates the progress of the underlying MapReduce job. When the job completes, the business analyst can inspect the results and tailor the collection further, if desired.

Figure 11 illustrates a sample sheet containing data similar to what you might find on a social media site.

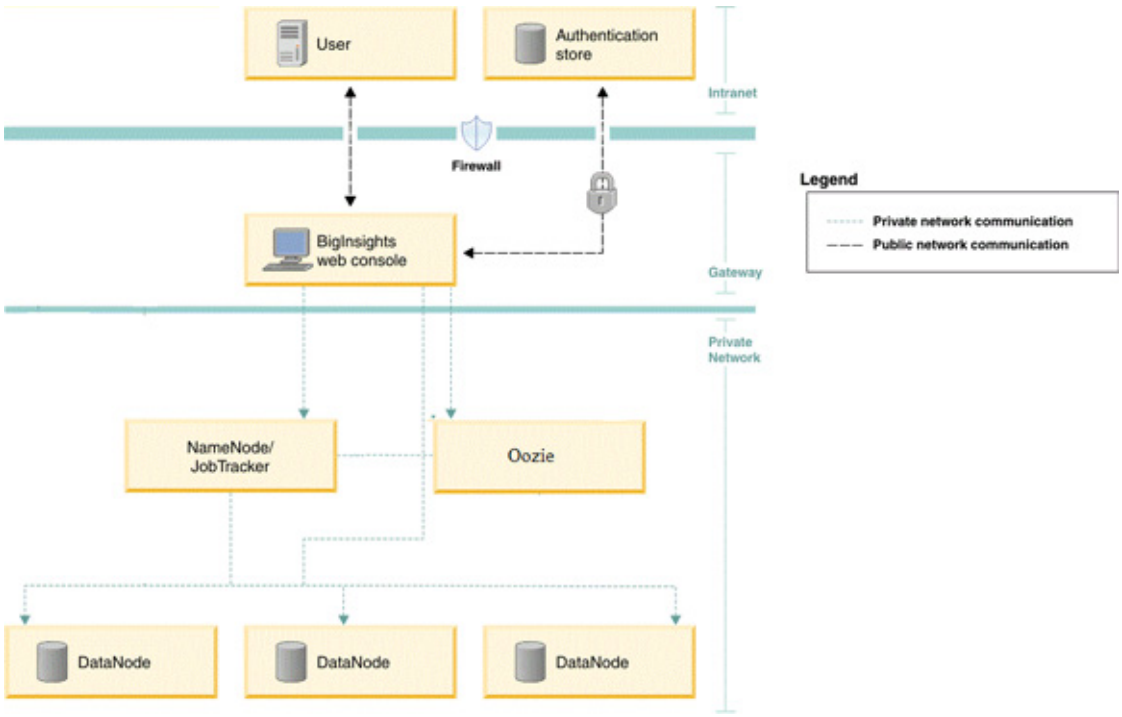
Figure 11. Analyzing and manipulating BigInsights data using a spreadsheet-style tool



Understanding security

BigInsights provides various enterprise security features that enable businesses to secure their cluster and data from unauthorized access. In a typical enterprise configuration, all BigInsights cluster servers are secured behind a firewall and connected over a closed network, with the web console serving as the gateway into the cluster. As shown in Figure 12, such secure configurations allow for unrestricted communication between cluster servers, while all ports are closed and rendered inaccessible from outside the cluster. The port serving the web console (by default, port 8080 for HTTP and port 8443 for HTTPS) is the only port that remains open to accept incoming communications. The BigInsights web console provides a reverse proxy feature that dynamically reroutes all HTTP traffic for the cluster through this single HTTP(S) port. The reverse-proxy function can be accessed via the Access Secure Cluster Servers link shown in Figure 4.

Figure 12. BigInsights secure reference architecture



The installer supports automatic setup and configuration for HTTP and HTTPS configurations.

Authentication

Authentication refers to the process of confirming that a user is indeed who he claims to be. The BigInsights web console supports three password-based authentication schemes. The recommended authentication setting for enterprise installations is LDAP. This approach enables you to configure the web console to perform authentication and group look-ups from an LDAP server. The web console can use LDAP or LDAPS (LDAP over SSL) protocols to communicate with the LDAP

store. The BigInsights installer provides detailed configuration options that enable companies to configure the LDAP server, the communication protocol, the LDAP subtree for user and group lookups, etc.

Flat file authentication allows administrators to configure the web console to look up a set of two properties files for user authentication and groups, respectively. These property files are located in the $BIGINSIGHTS_HOME/console/conf/security directory. The biginsights_user.properties file consists of entries of the form `user=password` and serves as the authentication store. The biginsights_group.properties file consists of entries of the form `group=user1,user2,...` and serves as the group lookup repository. Passwords stored in the user properties file can be secured using MD5 or SHA1 encryption with a hex or Base64 encoding. The flat-file authentication option is commonly used for product demonstrations or preproduction setups where the security provided by a file-based authentication store suffices.

By default, the BigInsights web console is installed without any authentication, which means that users can access all console functionality without entering any user ID or password. (This is consistent with Apache Hadoop 0.20.2.) Although this option is sufficient for exploring the web console functionality, it is not a suitable for enterprise installations.

Authorization

BigInsights supports role-based access control for all file system access, cluster administration tasks, application lifecycle management, and execution of applications published in the catalog. During installation, enterprise users and groups can be mapped to the four BigInsights roles with predefined privileges:

1. The BigInsights system administrator can perform all system administration tasks, such as monitoring cluster health and adding, removing, starting, and stopping nodes.
2. The BigInsights data administrator is authorized to perform all data administration tasks, such as creating directories, running Hadoop file system commands, and uploading, deleting, downloading, and viewing files.
3. The BigInsights application administrator can perform all application administration tasks, such as publishing and deleting an application, deploying and un-deploying an application to a cluster, configuring the icons, applying application descriptions, changing the runtime libraries and categories of an application, and assigning permissions of an application to a group.
4. The BigInsights user is possibly the most commonly granted role to cluster users who perform non-administrative tasks. Users can run applications that he has permission to run and view the results, data, and cluster health.

A simple command-line utility enables administrators to update the role mappings and keep them up to date post-installation. The utility, located at

$BIGINSIGHTS_HOME/console/bin/refresh_security_config.sh, reads the contents of the install XML file from $BIGINSIGHTS_HOME/conf/install.xml and redeploys the web console based on the current settings.

Acknowledgements

The authors would like to thank colleagues who worked on this technology and contributed ideas to this article. In addition, the authors welcome Yu Gao's new focus on BigInsights security, a previous specialty of Priya Baliga. Yu Gao can be reached at ygao@us.ibm.com.

Summary

The BigInsights web console provides tools for administering your cluster, launching applications and monitoring their status, working with your distributed file system, and analyzing data using a spreadsheet-style tool. This article introduced you to many important aspects of the web console in an effort to help you get off to a quick start with your BigInsights projects.

Resources

Learn

- Watch the Big Data: InfoSphere BigInsights v1.3 console demo to see Steve Brodsky demonstrating many of the technologies discussed in this article.
- Read "Understanding InfoSphere BigInsights" to learn more about the product's architecture and underlying technologies.
- Watch Big Data: Frequently Asked Questions for IBM InfoSphere BigInsights to listen to Cindy Saracco discuss some of the frequently asked questions about IBM's Big Data platform and InfoSphere BigInsights.
- Visit the BigInsights Technical Enablement wiki for links to technical materials, demos, training courses, news items, and more.
- Check out BigData University for free courses on Hadoop and big data.
- Refer to the IBM InfoSphere BigInsights Information Center for documentation about the product.
- Order a copy of *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* for details on two of IBM's key big data technologies.
- 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.
- Now you can use DB2 for free. Download DB2 Express-C, a no-charge version of DB2 Express Edition for the community that offers the same core data features as DB2 Express Edition and provides a solid base to build and deploy applications.

Discuss

- Participate in the discussion forum for this content.
- Check out the developerWorks blogs and get involved in the developerWorks community.

About the authors

Cynthia M. Saracco

Cynthia M. Saracco works on database management and XML technologies at IBM's Silicon Valley Lab. She has co-authored three books and taught university-level courses on various software technologies.

Priya Baliga

Priya Baliga has served as an advisory software engineer and technical lead at IBM's Silicon Valley Lab working on Big Data technologies. She began working with IBM in 2004, after acquiring her master's degree in computer science. She has served in various database development roles, including leadership roles in database management and security. She has patents and publications in various aspects of information security and management.

Stephen A. Brodsky

Stephen A. Brodsky is a technical executive and Distinguished Engineer for IBM Big Data initiatives at the IBM Silicon Valley Laboratory. Big Data is the strategic integration of large-scale information processing, including Hadoop map-reduce, streams, database, web servers, indexing, analytics, ETL, modeling, and traceability for structured, semi-structured, and unstructured information. Previously, he led the architecture for the Optim Data Studio product line and pureQuery, and was a member of the architecture team for DB2 pureXML, Rational Application Developer (RAD), and WebSphere. Brodsky holds doctoral and master's degrees in electrical and computer engineering, and a joint bachelor's degree in applied mathematics and biochemistry and cell biology. He has filed more than 40 patent applications.