Chain and link predefined sequence applications using InfoSphere BigInsights

Automating deployment through chained code

Timothy Landers (landertr@universalinet.com)  
Consultant  
Universalinet.com, LLC  
17 June 2014

By chaining and linking sequences of code, software developers can now create custom applications using InfoSphere® BigInsights™, organize data sources using BigSheets, enhance Apache Hadoop indexing with BigIndex, and even schedule automated MapReduce jobs using the InfoSphere BigInsights scheduler. Discover how your organization can gain value-added productivity without having to purchase additional software.

IT software developers purchase software applications with the expectation of getting multiple uses from the same product. InfoSphere BigInsights provides a means of creating, chaining, and linking predefined sequences of code developed within InfoSphere BigInsights to create custom software applications without adding code.

Customized applications address user and system requirements, and business objectives by delivering technology solutions that solve problems, enhance processes, and improve operational performance. Custom applications enable you to reuse code by applying existing technology to new uses to create value and revenue. From custom features such as remembering login information to saving repeated searches, settings, or user preferences, customized software applications are an excellent way to add business value.

User and system requirements

Customizations take into account specific user requirements. To develop an InfoSphere BigInsights application, developers must take into account how the new functionality affects automated operations, such as continual processes and procedures. Functional and nonfunctional requirements development ensures that the resulting applications work for system and human interaction processes.
BigInsights scheduler, BigIndex, and BigSheets

A foundational set of tools is built into InfoSphere BigInsights to help analyze data and integrate the analysis with other tools and functions. These tools enable software developers and administrators to enhance their InfoSphere environments by producing persistent results-driven automations for insightful solutions to business and technology requirements.

InfoSphere BigInsights is compatible with open source software solutions. The benefits of the open source framework are obvious: faster performance and greater scalability. Administrators can use the InfoSphere BigInsights scheduler for job allocation and can effectively prioritize jobs by the time required to run them.

InfoSphere BigInsights provides a significant level of security through LDAP authentication. Reverse proxy support and LDAP authentication actively help to prevent unauthorized access to the InfoSphere BigInsights console. A welcome addition to any administrative portal, the security features also include compressed data using IBM LZO-based compression technology and adaptive runtime jobs for Jaql.

The open source Apache ZooKeeper project is central to the infrastructure of InfoSphere BigInsights. It cross-synchronizes services to the cluster environment.

Application sequences

Three tools offer a great deal of agility in creating sequences of customized applications:

- **InfoSphere BigInsights scheduler** — Helps assign MapReduce jobs to a workflow. Working in conjunction with the Apache Hadoop Fair Scheduler, the InfoSphere BigInsights scheduler virtually guarantees that fair and equal cluster resources are provided to every MapReduce job. The fair aspect of the scheduler is the first-in, first-out approach to processing job queues.

- **BigIndex** — Transfers native capabilities to Hadoop indexing. This feature initializes Hadoop full-text searches using index scanning and querying. The BigIndex tool is a module — a component of a workflow — you can integrate as a component of other workflows to augment, enhance, replace, or add one or more software application functions.

- **BigSheets** — Shows (or visualizes) clusters so users can discover and analyze data in various clusters using queries and algorithms (built-in macros). Because it's web-based, BigSheets can integrate unstructured web information at massive scales, even beyond petabytes of data.

Jaql is the query language that ties it all together. Imagine needing to generate reports in answer to detailed questions from IT auditors. Your environment consists of huge amounts of big data, the majority of which is unstructured. Typically, time is of the essence in mission-critical and high-profile situations, such as IT auditing. Jaql provides powerful oversight of structured and unstructured data. InfoSphere BigInsights comes with predefined Jaql modules that serve as components for Apache Lucene indices and that work with Apache HBase (the Hadoop database) and IBM Netezza appliances for data warehouses. Jaql is resilient, fast, and accurate — accomplishing predictive analytics at real-time and near-real-time speeds. Meeting the needs of recent data innovations, the Jaql modules support text analytics as part of the language’s workflow and software component integration.
Data warehousing solutions

You can implement InfoSphere BigInsights as a data warehousing source. You can combine data from various sources, aggregate it, and assemble it into an InfoSphere BigInsights data warehousing solution. This approach enables you to add plugins and custom applications that automate and orchestrate redundant, routine processes. The applications and plugins facilitate the processing of structured and unstructured data to gain new insights from new data patterns. Analysis is heightened, reporting and querying are optimized, and business intelligence finds its central source for data retrieval.

Process to create custom applications

You can download InfoSphere BigInsights Quick Start Edition at no cost. With this edition, you can quickly extend the ability of applications such as Hadoop. From Big SQL to BigSheets, you can use InfoSphere BigInsights to easily build custom applications, without writing code.

InfoSphere BigInsights Quick Start Edition

InfoSphere BigInsights Quick Start Edition is a complimentary, downloadable version of InfoSphere BigInsights, IBM's Hadoop-based offering. Using Quick Start Edition, you can try out the features that IBM has built to extend the value of open source Hadoop, like Big SQL, text analytics, and BigSheets. Download BigInsights Quick Start Edition.

InfoSphere BigInsights enables you to use predefined InfoSphere BigInsights tools to create custom software application components, which are snippets of code that perform a specific task and have their own start and ending code. You can link InfoSphere BigInsights components to each other or to other software components. By snipping or dissecting the component or by referring to or calling the component into existing code sequences, you can integrate the components with other functions. Then you can chain the entire sequence of linked components to form a software application.

Administrators, users, and software developers all play a role in the creation of components. For example, administrators install and manage files and clusters, users use the custom applications to search for big data and process calculations and analyses, and software developers chain and link components into custom applications. Take a look at the process.

Pre-built modules offer easy-to-use functions to users. Extending BigSheets support to user-defined functions makes it possible to publish to an InfoSphere BigInsights server via the Eclipse IDE. To support software analysts, configuration managers, programmers, and nontechnical staff, BigSheets offers a spreadsheet for data analyses. The spreadsheet enables you to validate that applications were created to your specifications.

The goal of using BigSheets is to organize data sources so they're easier to scrutinize, examine, and analyze. BigSheets also supports JSON data, comma-delimited and comma-separated data, and other mainstream data formats. BigSheets supports big data and Hadoop capabilities and enables you to create plugins that run in the background to gather data. You can also define BigSheets collections and invoke them to create MapReduce jobs for data retrieval. Whether you're using macros or custom functions, BigSheets helps you customize automation and orchestrations for data handling.
As a first step, create an InfoSphere BigInsights server and associate it with the Eclipse IDE. In the figures that follow, notice that the information is organized into labeled fields and tabs for ease of use. The result of using the InfoSphere BigInsights console, the Eclipse IDE, and associated tools to develop applications is a compiled application, but no coding is required. Fill in the fields and click **Apply** and **Run** (to test) or **Close** (when finished).

When the application is published, you deploy the InfoSphere BigInsights function by using the InfoSphere BigInsights console.

**Figure 1. InfoSphere BigInsights server panel**

Now that the InfoSphere BigInsights server is created, select **Create New Application**. Name the application, enter a description, and select a category in which the function is to reside. As shown below, the application retrieves Twitter messages about IBM Watson™ and uses the associated icon.
Figure 2. Specify an InfoSphere BigInsights application to publish

Using the options in the InfoSphere BigInsights console shown in Figure 3, you can create or import queries from Jaql, create module-driven application components, and link and chain sequences of modules you have already created. The customizable dashboards show the status of the applications, data services, and clusters and the status of Hadoop components and various system attributes. Of particular interest is the list of quick links that make importing sample and module applications easy to do.

From the console, you can deploy and run applications that generate specific data and that work with BigSheets to visualize the data. You gain the power to build, deploy, and run custom applications that offer time savings, process improvements, and accelerator capabilities. Consolidating massive amounts of big data requires automations that make the process shorter and easier to perform. Recursive procedures benefit the most from these custom applications, adding accuracy, timeliness of delivery, and efficiency. In addition, function sheets operate as workbooks in which to store the list of applications, categorized as you specify.
Jaql, for example, needs only the paths to search to perform the ad-hoc queries you specify. The Applications tab, shown below, offers various options for defining, deploying, and "undeploying" the application types (for example, database import and export modules, a board reader module, and various ad-hoc modules for querying).

Figure 4. The console Applications tab

In the Edit configuration and launch window, shown below, you can edit a configuration prior to launching the Jaql script. The Arguments tab applies to the configuration. You can add arguments that further constrain the application's functions.
You can also use the deployment process to select the workflow that is to receive the new function. You can link individual or chained custom applications to a workflow to enhance the functions the workflow can perform.
After deployment, the application can be accessed by users.

In a similar manner, you can use InfoSphere BigInsights Quick Start Edition to perform the most popular InfoSphere BigInsights application development tasks, but the window includes a more condensed set of fields, settings, and parameters. This edition also shows cluster status and provides a dashboard, a view of the list of files, an Applications tab, an Application Status tab, and a BigSheets tab. Each tab is a shortcut to the most commonly used features and functions of InfoSphere BigInsights. As a result, you and your administrators can use the advanced module and component tools to generate custom applications.
The IDE creates a JAR file and transfers it to the InfoSphere BigInsights server when the application publishing process is complete.

Predefined sample applications in InfoSphere BigInsights

Sample applications include ad-hoc queries using Apache Hive, Jaql, and Apache Pig, and ad-hoc scripting using R. Additional sample applications perform downloads, create subsets of data, import and export data, copy data to or from the Hadoop Distributed File System, and make copies of files and directories. Sample applications are available for sorting, crawling the web, retrieving information, performing word counts, and more.
Conclusion

This article explains how to set up, configure, and test custom applications using InfoSphere BigInsights. When you configure the application, you can integrate plugins, specify types of data to target, restrict the environment, and define various interoperable paths to disparate data. The ability to customize applications based on a specific set of users produces new solutions that common business problems.
Resources

- Check out Persistent's take on InfoSphere BigInsights — Big Data Platforms for more information about application development and data automation.
- Check out the Cynthia M. Saracco interview by the ODBMS Industry Watch, Setting up a Big Data project, which discusses how to get started with a big data project.
- Read Implementing IBM InfoSphere BigInsights on IBM System x, by Mike Ebbers and others.
- Read "Big Data Reveals Big Insights," which covers mainstream big data problems, aims, and solutions for industry-standard environments.
- Stay current with developerWorks technical events and webcasts focused on a variety of IBM products and IT industry topics.
- Watch developerWorks on-demand demos ranging from product installation and setup demos for beginners to advanced functionality for experienced developers.
- Follow developerWorks on Twitter.
- Join the developerWorks community, a professional network and unified set of community tools for connecting, sharing, and collaborating.
About the author

Timothy Landers

Timothy Landers, a principal at Universalinet.com, LLC, is a practice lead in an independent consultancy. He has an MBA in technology management and is a Project Management Institute-certified Project Management Professional with more than 15 years in increasingly more-responsible roles within the IT field. He has written more than 28 technical courses for corporate training, vocational training, and higher education, as well as new product manuals, professional certification exams, and commercial sales catalogs (such as SkillSoft).

© Copyright IBM Corporation 2014
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)