

Informatica HParser

Simplify Hadoop Logic Development

Arun C. Murthy

Co-founder

MapReduce project in Apache Hadoop
Hortonworks

“As we advance the Apache Hadoop framework for storing, processing and analyzing big data, Informatica HParser’s approach in enabling parallel parsing across the Hadoop clusters available inside MapReduce is a unique addition to the Apache Hadoop community. Informatica HParser on the Hortonworks distribution uniquely takes advantage of the Apache Hadoop programming framework and is lineally scalable.”

BENEFITS

- Easily manage multiple industry standards, various binary documents, and complex hierarchical data
- Develop with ease using an intuitive UI that allows viewing data samples in their original and text formats
- Parse any data format from within Hadoop, using a single transformation engine accessible for the MapReduce developer in a simple call
- Rapidly transform structured and semi-structured formats using a visual development environment

Universal Data Handler for Hadoop

Informatica HParser enables access to the most difficult data and file formats in Hadoop, reducing the time and cost of developing data handlers by 70 percent. Utilizing the Data Transformation Studio, a visual example-based parsing environment, the user can easily define how to transform complex data into flattened, usable formats to be leveraged by Hive, PIG and further MapReduce processing. Greatly simplifying Hadoop logic development, it employs an engine within a single call to parse and handle any format, allowing the developer to work with a single data abstraction layer.

Using Informatica HParser, organizations can:

- Define complex data handlers and parsers in a visual example-based environment, with out-of-the-box support for logs, built-in support for binary formats, and packaged transformations for industry formats
- Immediately deploy and reuse transformations in Hadoop MapReduce
- Normalize multiple data formats to a single data abstraction, using a single transformation

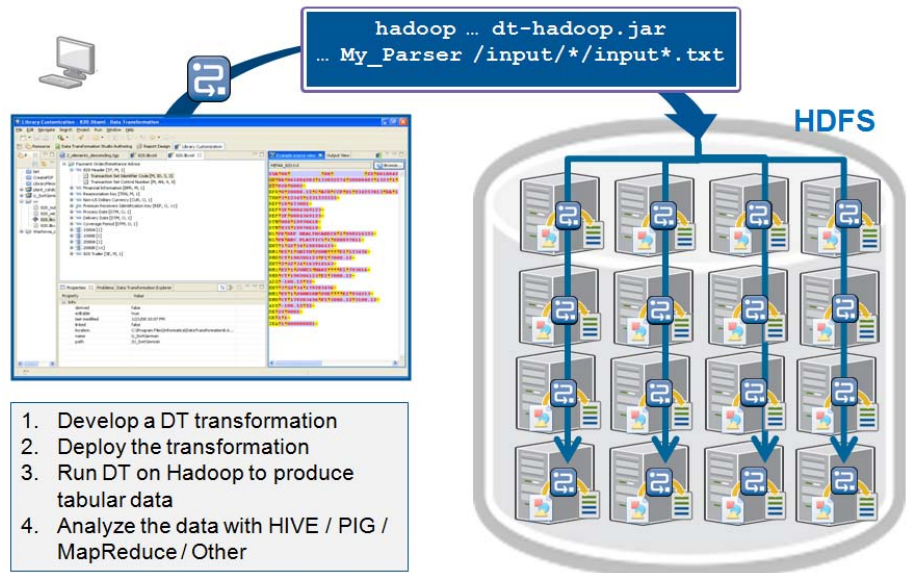


Figure 1. How HParser operates to parse and handle data in Hadoop.

Key Features

Broadest support for data formats

Wide support for multiple industry standards. Processing industry specific data standards and formats – such as EDI for manufacturing, SWIFT, NACHA and SEPA for Payments, ACORD for Insurance, ASN.1 for telco, HL7 for healthcare and more – poses a particular challenge. These standards, typically defined by industry groups or government organizations, are continually evolving and introducing new and enhanced formats. Most of these standards have at least one new version per year, requiring any multiyear Big Data analytics initiative to support multiple versions and variations. With our broad set of libraries, versions, and messages, Informatica HParser eliminates this problem, providing regular updates of new and existing standards soon after their release by industry and government organizations. With these updates, the current process is able to support new formats as they become available.

Unique support for binary documents: Word, Excel, PDF. Organizations store huge amounts of data in documents, such as legal files and contracts in Word and PDF, and financial reports and forecasts in Excel. Informatica HParser offers unique out-of-the-box support for these binary documents, allowing users to process and extract relevant data from them into Hadoop.

Robust support for hierarchical data. Formats such as XML and JSON increase the complexity of hierarchical data. The ability to effectively process data from a deep hierarchy and to support advanced schema and structures is required to successfully process the complex data in these formats. Informatica HParser features native support for XML and JSON as well as an optimized approach to extracting data from hierarchical structures.

Support for logs. Utilizing a patented specifications-driven transformation engine, Informatica HParser facilitates the defining of logs specifications, including hierarchical, delimited, and positional logs. These specifications can also be leveraged to parse and extract data from logs, including web logs, call detail records logs, mainframe logs, and proprietary logs.

Ease of development with intuitive UI

The unique example-driven transformation capability of Informatica HParser dramatically increases productivity, as compared with more traditional data handling and parsing approaches such as Java, Perl, and XSLT mappers. Whether the source is a PDF file, Excel spreadsheet, COBOL copybook record, or a Bloomberg, SWIFT, or ASN.1 message, the HParser user interface enables viewing a data sample in its original and text format. The example-based approach allows the continuous development of the parser or data handler, and provides instant feedback without the need to compile and deploy.

Integration and deployment in MapReduce

Informatica Data Transformation provides universal deployment across different environments and seamless integration within a wide variety of software infrastructures across all topologies. It plugs directly into any SOA, EAI, B2B, or data integration platforms. Informatica HParser further enhances that deployment flexibility with built-in integration to Hadoop MapReduce. The HParser engine is accessible for the MapReduce developer in a simple call, enabling the parsing of any data format inside Hadoop.

Rapid data abstraction for the Hadoop (MapReduce) developer

Advanced Big Data analytics scenarios depend on the ability to process data from multiple sources. Informatica HParser provides a visual development environment to transform these structured and semi-structured formats rapidly into a usable, canonical, and flattened format. The use of HParser as a single transformation engine, instead of multiple coded data handlers, allows the MapReduce developer to develop a single program agnostic to the data variation.

Learn More

Learn more about the Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the U.S.).

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to gain a competitive advantage with timely, relevant and trustworthy data for their top business imperatives. Worldwide, over 4,440 enterprises depend on Informatica for data integration, data quality and big data solutions to access, integrate and trust their information assets residing on-premise and in the Cloud. For more information, call +1 650-385-5000 (1-800-653-3871 in the U.S.), or visit www.informatica.com. Connect with Informatica at <http://www.facebook.com/InformaticaCorporation>, <http://www.linkedin.com/company/informatica> and <http://twitter.com/InformaticaCorp>.



Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871 www.informatica.com