unifi
Enterprise-class self-service data integration

WHITEPAPER

Unifi OneParse™ Automated Parsing of Semi-Structured Data

By Ayush Parashar, Ravikiran Krishnan, Sean Keenan, Sudeep Sarkar, Deepak Chandrasekar

Summary

Providing simple solutions that democratize access to data is at the heart of generating value for the enterprise. However, most enterprises have a huge amount of data that is semi-structured or un-structured. It makes it hard for non-technical user to work on such data. Previously, manual parsing tools, employed by programmers, were the only way to solve the problem of importing and using semi-structure datasets. These datasets are generally very wide and varying in nature. Manual clean-up of such datasets can be very labor intensive and time consuming. Also, there are other challenges such as identifying missing values and reconstructing them.

To provide a holistic solution to this problem, Unifi Software has developed a proprietary approach to parse the semi-structured data automatically and represent the data in a flattened, tabular structure that is easy to consume by the end-user. Unifi goes one step further, in identifying missing columns and bad rows. Additionally, datasets can be enhanced through a simple, intuitive user interface.

Figure 1: Example of semi-structured data and a structured representation of it.

Semi-structured Data

Semi-structured data is prevalent in enterprises in the form of weblogs, data gener-ated from sensors, application logs, etc. While structured data is comprised of rows and columns (or structured through a known schema) with simplistic use of a single character column delimiter, the semi-structured data is different in the following ways:

```
[unifi@487444b4275b docker]$ cat weblog.log
133.43.96.45 - - [01/Aug/1995:00:00:22 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /shuttle/missions/sts-69/sts-69-patch-small.gif HTTP/1.0" 200 8083
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
133.43.96.45 - - [01/Aug/1995:00:00:25 -0400] "GET /history/apollo/images/apollo-logo1.gif HTTP/1.0" 200 1173
133.43.96.45 - - [01/Aug/1995:00:00:46 -0400] "GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0" 200 6168
133.43.96.45 - - [01/Aug/1995:00:00:22 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /shuttle/missions/sts-69/sts-69-patch-small.gif HTTP/1.0" 200 8083
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
133.43.96.45 - - [01/Aug/1995:00:00:25 -0400] "GET /history/apollo/images/apollo-logo1.gif HTTP/1.0" 200 1173
133.43.96.45 - - [01/Aug/1995:00:00:46 -0400] "GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0" 200 6168
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /images/ksclogosmall.gif HTTP/1.0" 200 3635
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /shuttle/resources/orbiters/orbiters-logo.gif HTTP/1.0" 200 1932
133.43.96.45 - - [01/Aug/1995:00:00:22 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /shuttle/missions/sts-69/sts-69-patch-small.gif HTTP/1.0" 200 8083
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
133.43.96.45 - - [01/Aug/1995:00:00:25 -0400] "GET /history/apollo/images/apollo-logo1.gif HTTP/1.0" 200 1173
133.43.96.45 - - [01/Aug/1995:00:00:46 -0400] "GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0" 200 6168
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /images/ksclogosmall.gif HTTP/1.0" 200 3635
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /shuttle/resources/orbiters/orbiters-logo.gif HTTP/1.0" 200 1932
133.43.96.45 - - [01/Aug/1995:00:00:22 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /shuttle/missions/sts-69/sts-69-patch-small.gif HTTP/1.0" 200 8083
133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
133.43.96.45 - - [01/Aug/1995:00:00:25 -0400] "GET /history/apollo/images/apollo-logo1.gif HTTP/1.0" 200 1173
133.43.96.45 - - [01/Aug/1995:00:00:46 -0400] "GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0" 200 6168
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /images/ksclogosmall.gif HTTP/1.0" 200 3635
133.43.96.45 - - [01/Aug/1995:00:00:51 -0400] "GET /shuttle/resources/orbiters/orbiters-logo.gif HTTP/1.0" 200 1932
[unifi@487444b4275b docker]$
```

Figure 2: Sample semi-structured log data.

- Each column is delimited by neighboring columns with varying delimiters
- Each delimiter can be single or multi-character (single/ multi byte). The charac-ters that form a delimiter can be different from each other.
- Visually, it is possible to figure out delimiters. A regular parser is not useful for such data.

Currently, to make use of semi-structured data, programmers have to create handwritten code or use a tool to manually convert it to a clear representation that can be used by upstream applications for creating analysis and to derive insight. This proc-ess is time consuming and reduces the access of such data to number of enterprise users. While we feel that it is important to give a simple and easy user interface that is usable by any enterprise user, automating the parsing process is very important for democratizing the use of semi-structured data across the enterprise.

Unifi Experience

At Unifi, our focus is on simplifying the experience around data integration and delivering a user experience that permits greater adoption of data analytics by line-of-business managers across the enterprise. One aspect of the Unifi end-to-end data integration tool set, delivers a dramatically simplified process for importing and employing semi-structured data. End users just browse a simple file-system interface to select a file. The file is automatically recognized and parsed by our patent-pending OneParse technology. The user is shown a simplistic tabular interface through which they can further enhance the structure or use it for transformation and other integration tasks.

OneParse Technology

At the heart of the simple experience is the patentpending Unifi technology OneParse. It is the superset algorithm that is responsible for multiple tasks related to data discovery including the following:

- Ascertain whether the data is structured or semistructured
- Parse the semi-structured data: Understand various delimiters, align column information, understand data-types for each of the columns
- Establish any missing values for different columns
- Identify badly written rows
- Represent the parsed semi-structured data and the missing value information for further data integration tasks
- Convert the representation into appropriate Hadoop execution elements for parsing the data at huge scale that takes care of missing data and error handling for badly formatted rows

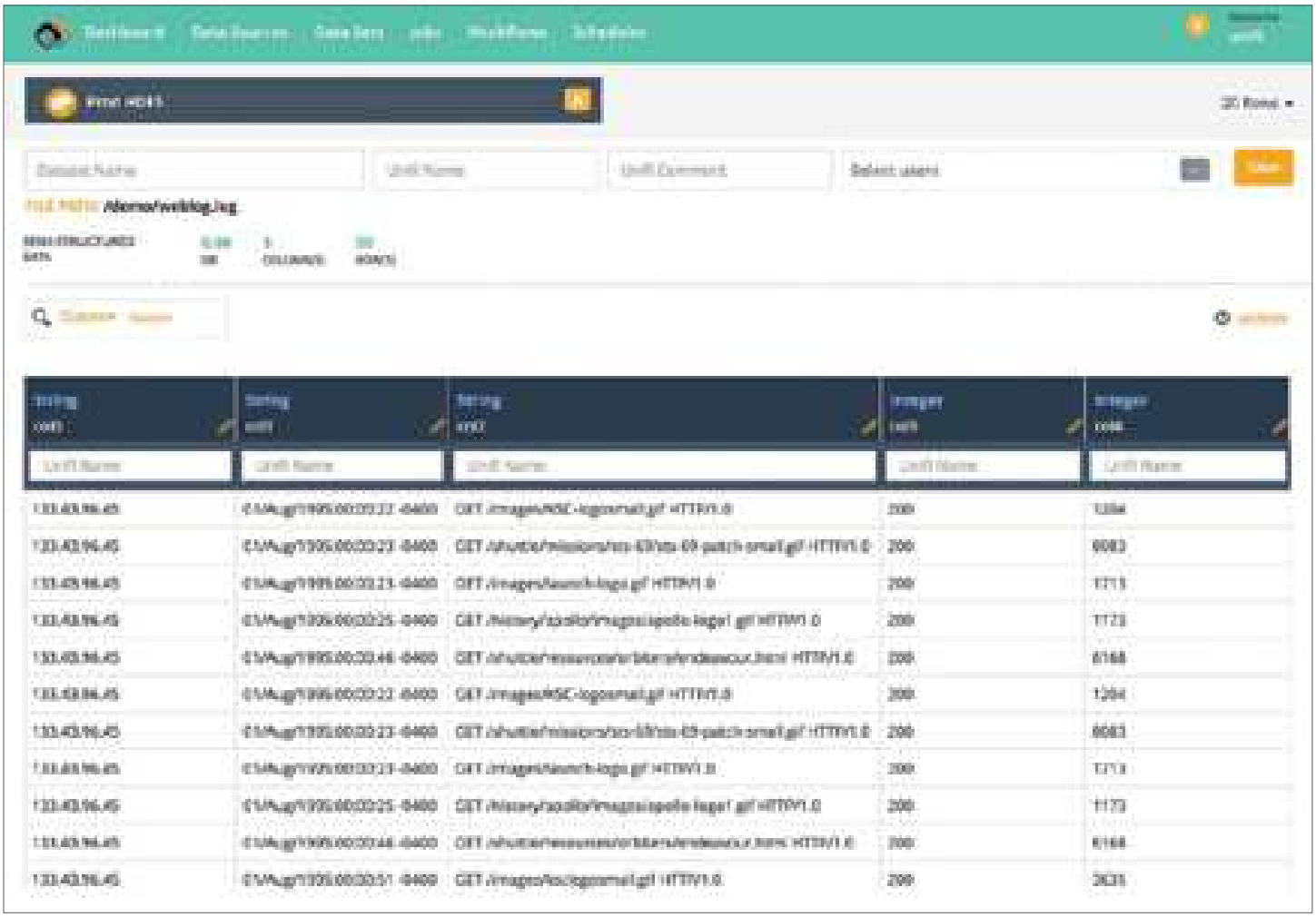

Figure 3: Structured representation of log data in Figure 2 after being parsed through OneParse™ Algorithm.

OneParse uses a combination of algorithms such as “dynamic time warping” and “center start” methods to solve the complicated problem. These algorithms look at the alignment of multiple unstructured rows to find each occurrence of a common set of delimiters across the rows and establishing a structure consensus. The result is a set of delimiters (which can be a combination of different types) on which data can be made tabular. Other heuristic methods are used to optimize the computation for real-time response. If the data has a different combination of delimiters, then that is the clinching evidence, in order to identify bad rows and missing columns. The identification of the missing delimiters is used to reconstruct bad data into a structured data format that can be displayed in a tabular format.

The operator is completely hidden from the complexity of this technology and only sees the data displayed in a clean, tabular format.

User Interface

Unifi has created a very intuitive interface to further enhance the tabular structure output that is generated by OneParse. Through simple UI drag and drop elements a user can split certain columns on selected text attributes inside the column. Columns can also be merged or repetitive data inside the column can also be deleted. Knowledge of the regular expression or any form of grammar is not needed to manipulate the data. All functions performed on the user interface are translated into the appropriate Hadoop execution elements for parsing the data at a huge scale. The combination of OneParse and an intuitive user interface delivers powerful capabilities to the end user to achieve the desired data view. Figure 4 and 5 below show an example of how a column split can be achieved easily.

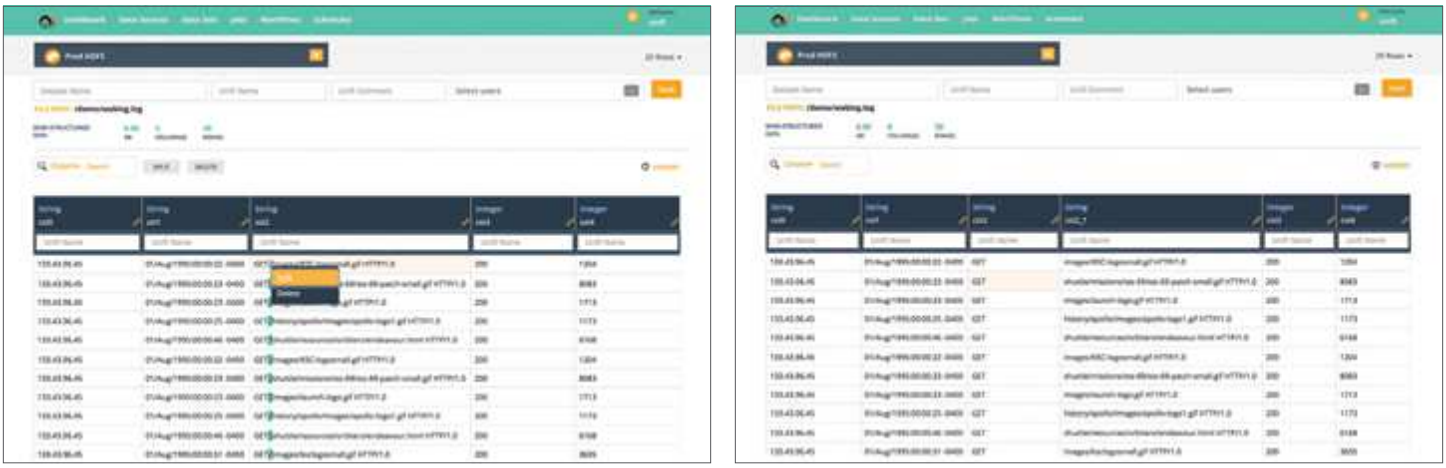

Figure 4: Splitting a column on a particular selection.

Conclusion

For an organization to take advantage of the data they own and to discover business insights in a timely manner, it is important to broaden the access of all kinds of data to business analysts across the enterprise. Through the introduction of the patent-pending OneParse technology, Unifi has been able to simplify the way complicated, semi-structured data can be consumed by business users.