SOLUTION SHEET

PERVASIVE

Data Movement and Transformation

Unstructured Data

Content Extraction

Harvesting Unstructured Data

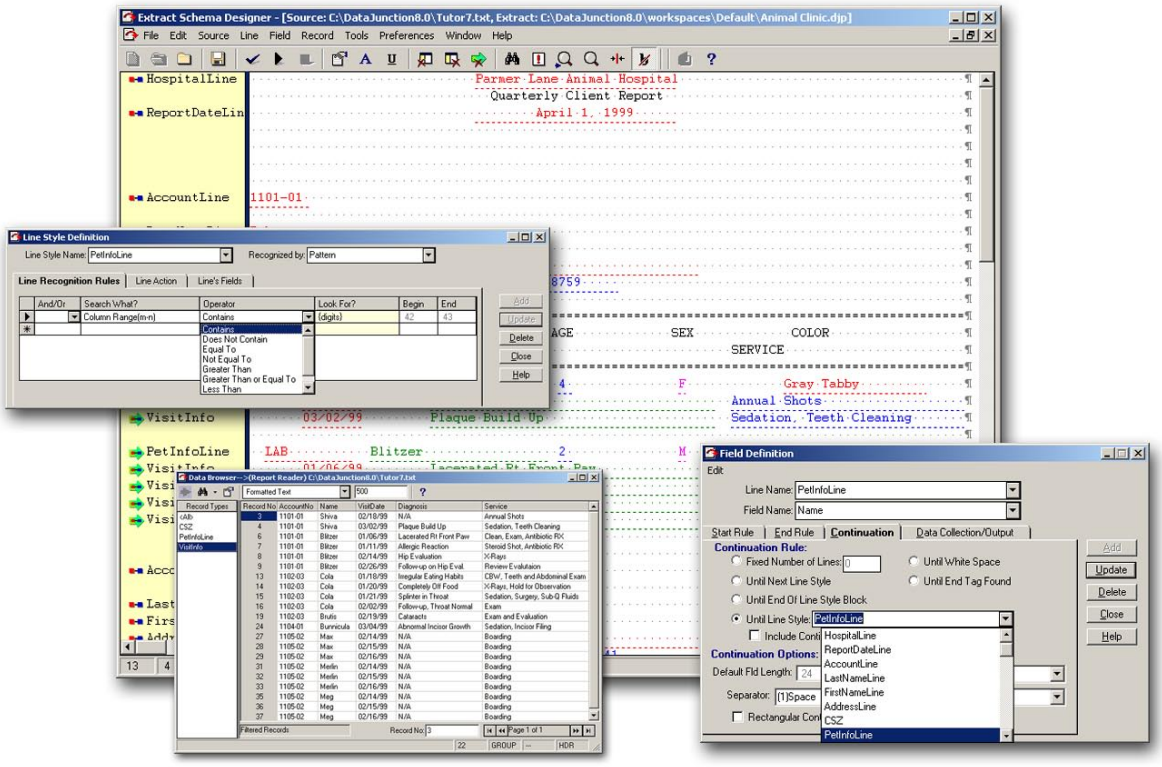
Even before the introduction of the World Wide Web, Gartner analysts estimated that more than 70% of the world's data was "lost" or "buried" in unstructured formats like documents. If you factor in the World Wide Web – the largest repository of unstructured data in history – you begin to realize that the vast majority of data in the world is accessible only as unstructured sources. Imagine the wealth of information and the strategic advantage you could gain by accessing and integrating this virtual gold mine of data!

Pervasive has positioned itself at the forefront of this problem by acquiring Content eXtraction Language (CXL) technology. This is the basis for the new Pervasive Internet Rapid Integration Services (djIRIS) launched in 2003. Furthermore, Pervasive Data Junction was recently selected by the readers and editors of Intelligent Enterprise for the 5th straight year as the #1 ETL tool for data movement and transformation, substantiating the company as the market leader for accessing and integrating data.

1. Pervasive Data Junction Extract Schema Designer

The key problem with unstructured data is a simple one, at least on paper: How do you tap in to and access the mountains of unstructured text data in the world? The good news is that once you're able to access these unstructured text formats, you have unfettered access to scores of valuable "structured" data – invoices, customer details, catalogs, addresses, etc. The bad news, of course, is that the data is "lost" in

Extract Schema Designer



>> continued

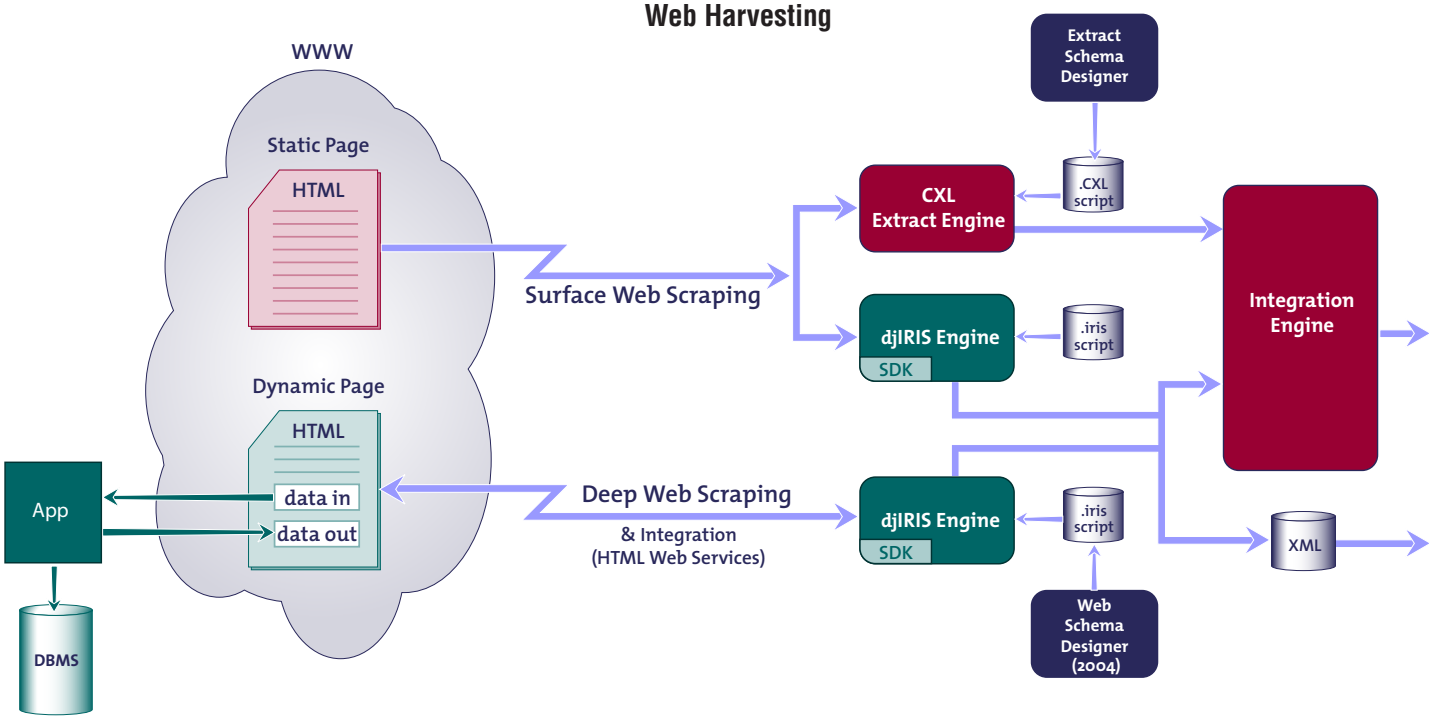
STRENGTHS:

- There is good news and bad news about unstructured data formats
- Non-programmers can quickly and easily access unstructured text file sources
- Integrate applications at the presentation layer and close the integration "loop"

ADDED BENEFITS:

- In today's Web-centric world, most data is only accessible in unstructured formats.
- Even before the explosion of the Web, Gartner analysts estimated that more than 70% of the world's data was "lost" or "buried" in unstructured formats like documents.
- Access and integrate unstructured data from a myriad of sources, tap a wealth of information and gain crucial strategic advantage.

PERVASIVE



unstructured formats with all the standard blemishes: floating fields, white space, page breaks, large text blobs, etc. The truth, however, is that these sources are not really unstructured. Rather, they are simply defined with structures that make them more readable by humans, not conventional, "machine-friendly" fixed and delimited text file readers. Consequently, the best way to access unstructured data is with a powerful pattern recognition language and graphical interface that allow extraction to occur in an automated fashion, but driven by text patterns – in other words, just like human readers.

The foundation for Pervasive's ability to extract structured data from unstructured text sources is the CXL Engine. The CXL Engine is a highly efficient line-oriented text manipulation and pattern recognition engine invented in, and built with, specific wiring to Pervasive Data Junction back-end high-speed Integration Engines. In just a few lines of code, one can extract perfect row and column "views" from streams of otherwise unreachable dirty text and simple HTML sources.

Since the real productivity gains in IT come from powerful end-user graphical interfaces, Pervasive Data Junction has built a powerful user interface on top of the CXL Engine. This enables non-programmer users to mark up the unstructured text file source quickly and easily and, in a matter of minutes, build an entire Extract Schema for even the most complicated unstructured text sources. The real beauty of the Extract Schema Designer is that it is in fact a "code generator" that is able to generate the CXL necessary to feed into Pervasive Data Junction unstructured text parsing engine. Consequently, users have the best of both worlds: they can quickly build and implement solutions using the Extract Schema Designer and, for those (admittedly rare) cases where additional performance or power is needed, they can fall back on the CXL language for the extra engineering horsepower they need. The Extract Schema Designer also includes visual debugging, source data "structured" viewers, add-on modules for PDFs and other document formats, as well as the ability to handle non-text binary and print characters. This complete infrastructure for accessing unstructured text sources has no rival in the industry, equipping users with a world-class toolset for unlocking the treasure of unstructured text data.

2. djIRIS - Internet Rapid Integration Services SDK

Pervasive Internet Rapid Integration Services SDK (djIRIS) solves the difficult yet essential problem of harvesting data from the greatest data source of all time: the World Wide Web. Again, this presents a good news/bad news scenario. The good news is that this data source represents an ocean of data of every conceivable kind from every conceivable source – and it resides literally at our fingertips. The bad news is that it is all locked away behind opaque HTML pages. And, in addition to the usual difficulties of navigating unstructured text-like HTML, there are

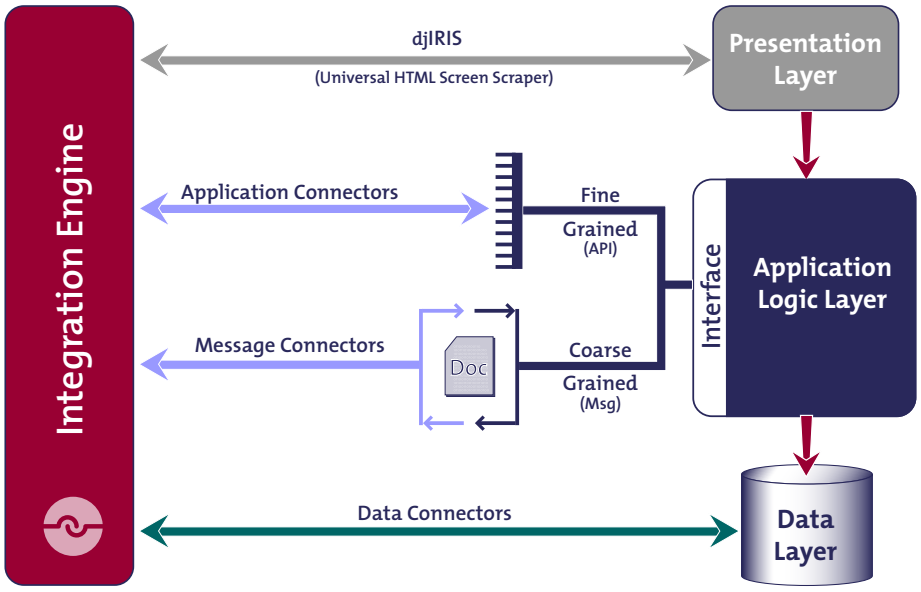
>> continued

the added barriers of HTTP and application-based authentication, as well as extremely complex navigation and looping scenarios to access the "leaf" pages targeted for harvesting.

To address this problem Pervasive has engineered, from scratch, a patent-pending djIRIS Engine that can act as a fully automated proxy browser, spoofing multiple browsers. djIRIS is a highly efficient and optimized language, based on the popular Java syntax, for controlling the behavior of an HTTP-based Web-browser agent. And, via a DOM-based XHTML infrastructure, djIRIS gives direct and automated control of a Web site to users. Guided by djIRIS scripts, or a very rich set of Java API calls, the djIRIS Engine intelligently traverses the World Wide Web and extracts useful structures of data of any shape or volume. The harvested data can then be delivered as XML, or fed directly to the high-speed Integration Engine for further downstream transformation and processing.

With the djIRIS Engine, two significant aspects of the Web are penetrated. First, there is the surface Web; the djIRIS Engine can harvest and deliver this data rather easily. This level consists of the 2,000,000,000+ pages of static HTML that we can think of as "generation 1" of the Web – i.e., Web pages Google can index. These HTML pages contain untold riches of data from a staggering variety of sources: internal or external, private or public. In many cases, these HTML surface pages are simply the most direct path – and sometimes the only path – to every kind of vital data needed for all sorts of business purposes (e.g., catalogs, documents, histories, etc.).

3 - Tier Application Integration



The "generation 2" is the "deep" web. This is the more interesting "hidden" web, the part that Google cannot touch, and is in fact not indexed at all. After some trial and error people quickly learned that it was impractical to simply dump the contents of their databases onto the surface web – the volume of the databases were too large and their content constantly changed. Consequently, the generation 2 Web became much more interactive (with CGI and other scripting alternatives), which helped to build more intelligent gateways or portal pages. These could ferry authenticated user requests from the front-end HTML page to the back-end database, and then return the results of the query in dynamically formatted HTML to the browser.

This generation 2 Web is rapidly dwarfing the generation 1 Web of static HTML pages. It is called the "deep Web" because it represents Web-based access to the real data treasures in the world – the thousands of huge databases that are otherwise locked behind firewalls, but are now integration-friendly via the magic of djIRIS. The scale of content access, aggregation and harvesting that this represents, via the unstructured medium of the World Wide Web, is truly staggering. And Pervasive Data Junction does this at a cost far below the pricey and less powerful alternatives offered by other vendors.

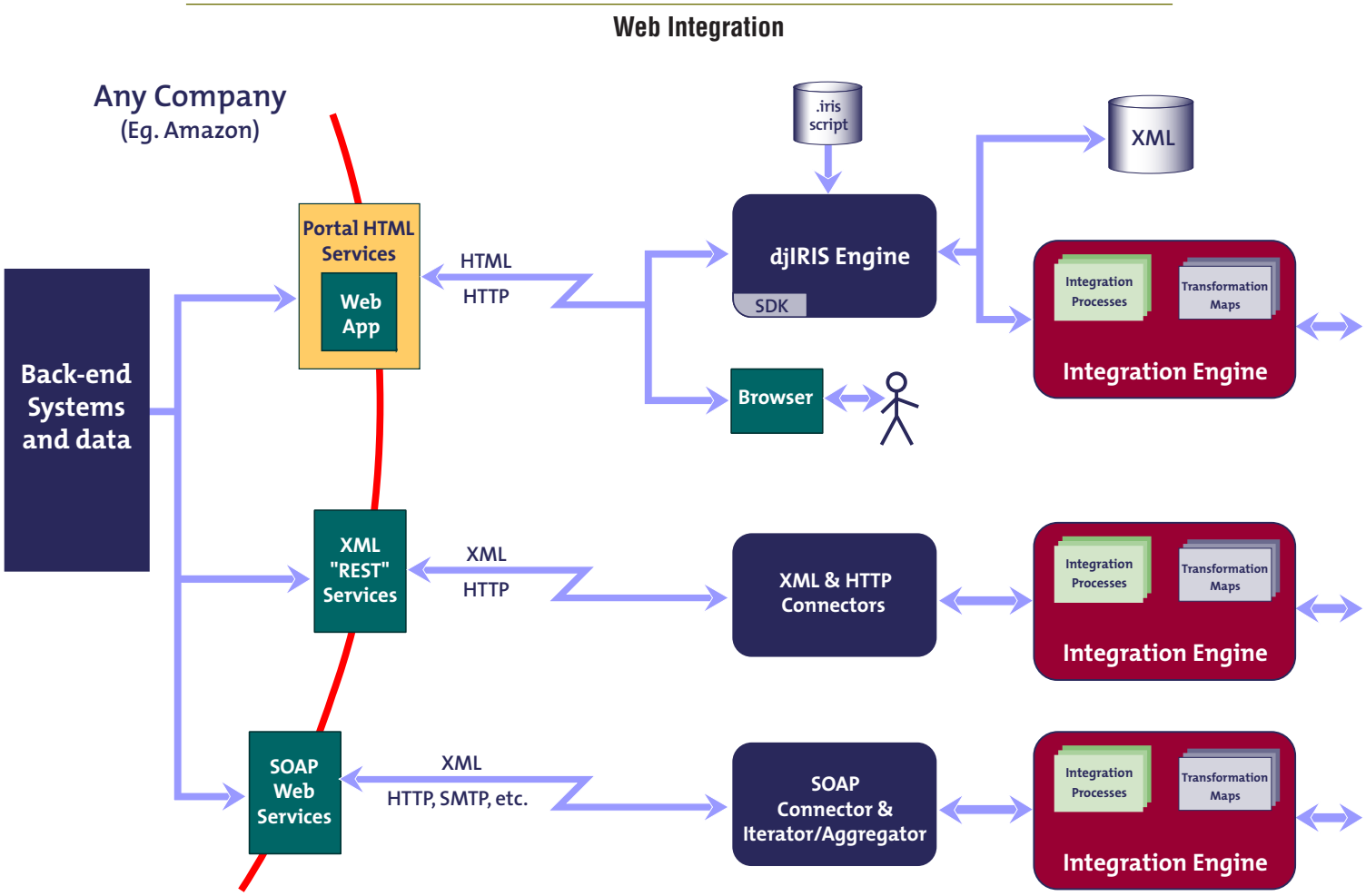
3. djIRIS Engine as Screen-scraping Integration Tool

Virtually all new application development is achieved with browser-based interfaces. djIRIS Engine, working directly at the browser interface level, and engineered to work bi-directionally (data entry and output) with all Web-based applications at the HTTP/HTML level, is a superb next-generation screen-scraping platform, leapfrogging all traditional legacy options (e.g., 3270, 5250, VT100 and PC). Above all, this gives users a powerful and dynamic new tool for integration projects.

While application integration at the application and logic layers – and occasionally at the data layer – is quite common, there are still scenarios when integration at the presentation layer is requisite. Since Pervasive Data Junction already delivers market-leading tools for integrating applications at the data and logic layers, having the djIRIS Engine included in your integration toolset for integrating applications at the presentation layer closes the integration "loop," providing you with the power of multi-level integration of every modern application in the world.

>> continued

PERVASIVE



4. djIRIS and HTML-based Web Services

With all the current hype surrounding XML-based Web services, it is easy to forget that there are already hundreds of thousands of "proto" Web services in existence, operating all over the world – both inside and outside the enterprise firewall. These Web services, engineered in HTML over HTTP, are often built with transaction semantics and are designed for human interaction via Web browsers. We encounter these types of Web services daily when we check stock quotes, use a search engine, or order merchandise from an online vendor. And when you consider the slow uptake of XML-based Web services, particularly at the B2Bi level (where there seem to be more Web services tools than Web services themselves), it becomes readily apparent that there are more new HTML-based Web services created every day than XML-based Web services created in a year. You can see why Pervasive Data Junction, in addition to supporting the relatively tiny and slow-growing market for XML-based Web services, is aggressively pursuing the rapidly growing number of HTML-based Web services in the world.

Unlike XML-based Web services where information exchange is automated by programs on both ends of the "exchange," HTML-based Web services occur when only one end of the exchange is automated. In a way, HTML-based Web services can be seen as an alternative form of XML-based Web services. Like XML-based Web services, the business logic of HTML-based Web services is exposed – but it is exposed to a human rather than to an automated program. Also like XML-based Web services, the interfaces for HTML-based Web services are well-defined; but in order to integrate with them in an automated fashion, an integration tool would have to simulate browser behavior – exactly the unique screen-scraping capability of the djIRIS Engine.

>> continued

SOLUTION SHEET

PERVASIVE

Conclusion

The powerful story cited above has played, and continues to play, a major role in Pervasive having the most widely deployed data integration tools in the world. In addition to our compelling technology, Pervasive Data Junction tools continue to enjoy the lowest TCO in the industry. This is not, however, simply because our up-front licensing costs are more attuned to today's economic reality; it is also because the ongoing running costs of our tools are much lower, and have a shorter life span, than custom code or our competitors' tools.

And as we continue to round out our extensive line of integration tools, you will see Pervasive Data Junction emerge as the only integration solution in the world with the home-grown technology and forward-looking vision to tackle all integration issues – from Web services (both XML and HTML) and B2Bi to EAI, ETL and data warehousing – on all major platforms, for enterprises of all sizes. Above all, with our nimble, high-speed, cost-effective tools, we enable all enterprises, with any integration challenge, to be effective in today's dynamic e-business world.

ABOUT PERVASIVE SOFTWARE

Pervasive Software is a leading global data management company powering the success of application developers by providing solutions that deliver the industry's best combination of performance, reliability and low administration costs. Pervasive's strength is evidenced by the size and diversity of its customer base, serving tens of thousands of customers with hundreds of thousands of end-users in nearly every vertical market around the world. Founded in 1994, Pervasive® sells its products into more than 150 countries and is based in Austin, Texas, with offices in Europe.

FOR MORE INFORMATION

- To learn more about Pervasive Software and our solutions, please visit **www.pervasive.com**.
- To reach the North American sales office, call **1.800.287.4383, extension 2**.
- For Latin, Central and South America, Australia and New Zealand, call **+1.512.231.6000**.
- In Europe, for Belgium, France, Germany, Italy, Luxembourg, The Netherlands, Spain, Sweden, Switzerland and the United Kingdom, call **+800.12.12.34.34**.
- For any other European, Middle Eastern, African or Asian countries (excluding Japan), call **+32.70.23.37.61**.
- For Japan, please call **+81.3.3293.5300**, or visit **www.pervasive.co.jp**.